DOCUMENT RESUME

ED 413 738 FL 024 759

AUTHOR	Rettig, Heike, Ed.
TITLE	Language Resources for Language Technology: Proceedings of the TELRI (Trans-European Language Resources Infrastructure) European Seminar (1st, Tihany, Hungary, September 15-16, 1995).
INSTITUTION	Institut fuer deutsche Sprache, Mannheim (Germany).
ISBN	ISBN-963-8461-99-3
PUB DATE	1995-00-00
NOTE	189p.; For individual articles, see FL 024 760-778. "In collaboration with Julia Pajzs and Gabor Kiss."
PUB TYPE	Collected Works - Proceedings (021)
EDRS PRICE	MF01/PC08 Plus Postage.
DESCRIPTORS	*Computational Linguistics; Computer Software; *Computer Software Development; Contrastive Linguistics; Czech; Data Processing; Dictionaries; *Discourse Analysis; Dutch; English; Foreign Countries; Information Technology; *Language Planning; Language Research; *Languages; Languages for Special Purposes; Linguistic Theory; Machine Translation; Morphology (Languages); Research Methodology; Russian; Shared Resources and Services; Slovenian; Spelling; Structural Analysis (Linguistics); Suprasegmentals; Uncommonly Taught Languages; Vocabulary
IDENTIFIERS	Speech Recognition

ABSTRACT

This proceedings contains papers from the first European seminar of the Trans-European Language Resources Infrastructure (TELRI) include: "Cooperation with Central and Eastern Europe in Language Engineering" (Poul Andersen); "Language Technology and Language Resources in China" (Feng Zhiwei); "Public Domain Generic Tools: An Overview" (Tomaz Erjavec); "The 'Terminology Market'" (Christian Galinski); "Lexical Resources and Their Application" (Martin Gellerstam); "Encoding Standards for Linguistic Corpora" (Nancy Ide); "Machine Translation: State of the Art, Trends and the User Perspective" (Steven Krauwer); "MULTEXT-EAST: Multilingual Text Tools and Corpora for Central and Eastern European Languages" (Erjavec, Ide, Vladimir Petkevic, Jean Veronis); "Speech Recognition: A General Overview" (Luis de Sopena); "Language Resources: The Foundations of a Pan-European Information Society" (Wolfgang Teubert); "Rail-Lex Slovenia--A Modern Railway Dictionary" (Primoz Jakopin); "A New Dutch Spelling Guide" (J. G. Kruyt, P. G. J. van Sterkenburg); "European Language Resources and the Treasury of the Computerised Russian Language Fund" (Elena Paskaleva); "HUMOR--A Morphological System for Corpus Analysis" (Gabor Proszeky); "CORDON--A Joint Venture Case Study" (Norbert Volz); "EVA--A Textual Data Processing Tool" (Jakopin); "On-line Access to Linguistically Annotated Text Corpora" (Kruyt, S. A. Raaijmakers, P. H. J. van der Kamp, R. J. van Strien); "Tagging a Highly Inflected Language" (Paskaleva, Bojanka Zaharieva); and "A Simple Czech and English Probabilistic Tagger: A Comparison" (Barbora Hladka, Jan Hajic). (MSE)

ED 413 738

ERI

LANGUAGE
STRUCTURE

PROCE F THE FIR
EUR EMINAR

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Norbert
Vol3
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ge Res s for
ge Tec y"

Tihany, Hu
ptember 15 an

Edited by Heik
boration with Júlia ss

BEST COPY AVAILABLE

PROCEEDINGS OF THE FIRST TELRI SEMINAR IN TIHANY

TELRI

TRANS-EUROPEAN LANGUAGE
RESOURCES INFRASTRUCTURE

PROCEEDINGS OF THE FIRST EUROPEAN SEMINAR

"Language Resources for Language Technology"

Tihany, Hungary
September 15 and 16, 1995

Edited by Heike Rettig
In Collaboration with Júlia Pajzs and Gábor Kiss

ISBN 963 8461 99 3

Contents

PREFACE 7
P. ANDERSEN
Cooperation with Central and Eastern Europe
in Language Engineering 9
FENG ZHIWEI
Language Technology and Language Resources in China 21
T. ERJAVEC
Public Domain Generic Tools: An Overview 37
C. GALINSKI
The 'Terminology Market' 49
M. GELLERSTAM
Lexical Resources and Their Application 57
N. IDE
Encoding Standards for Linguistic Corpora 65
S. KRAUWER
Machine Translation: State of the Art,
Trends and the User Perspective 79
T. ERJAVEC, N. IDE, V. PETKEVIČ, J. VÉRONIS
MULTEXT-EAST: Multilingual Text Tools and Corpora for
Central and Eastern European Languages 87
L. DE SOPEÑA
Speech Recognition: A General Overview 99
W. TEUBERT
Language Resources:
The Foundations of a Pan-European Information Society 105
P. JAKOPIN
Rail-lex Slovenia – A Modern Railway Dictionary 129
J.G. KRUYT, P. G. J. v. STERKENBURG
A New Dutch Spelling Guide 133
E. PASKALEVA
European Language Resources and the Treasury
of the Computerised Russian Language Fund 143

G. PRÓSZÉKY
HUMOR – A Morphological System for Corpus Analysis 149
N. VOLZ
CORDON – A Joint Venture Case Study 159
P. JAKOPIN
EVA – A Textual Data Processing Tool 169
J.G. KRUYT, S. A. RAAIJMAKERS, P. H. J. VAN DER KAMP,
R. J. VAN STRIEN
On-line Access to Linguistically Annotated Text Corpora
of Dutch via Internet 173
E. PASKALEVA, B. ZAHARIEVA
Tagging a Highly Inflected Language 179
B. HLADKA, J. HAJIČ
A Simple Czech and English Probabilistic Tagger: A Comparison . . 191

Preface

This book documents the presentations given at the first TELRI Seminar "Language Resources for Language Technology" held in Tihany, Hungary from September 15-16, 1995. TELRI (Trans-European Language Resources Infrastructure) is an international project funded by the European Commission. It brings together 22 institutions from 17 European countries. TELRI will set up a permanent network of leading national language and language technology centers and will pool existing language resources and generic software tools.

Language technology needs language resources such as corpora of spoken and written language, word lists, lexicons, machine readable dictionaries, and tools to extract linguistic knowledge to develop and optimise products. Since such kinds of language resources are often available in high quality in the domain of public research, one aim of the first TELRI Seminar "Language Resources for Language Technology" was to provide a platform where researchers in the field of corpus linguistics and natural language processing, lingware developers, and end-users of language resources can meet and discuss new ideas and possibilities for co-operation.

The TELRI Seminar provided information on the existing and emerging European language technology infrastructure, showed the state of the art in relevant fields of natural language processing, demonstrated joint venture projects between (public) research and (private) industry, and gave the opportunity to see computer demonstrations featuring resources, tools, and products in the field of language technology. All of these contributions came from researchers from the university domain (TELRI members, members from other international projects, members of research institutes) as well as from representatives of private companies. According to the challenges of a multilingual Europe, the seminar tried to establish an international perspective, reflected by the topics of presentations and the speakers and participants themselves who came from twenty-four different countries.

Since TELRI includes numerous members from Central and Eastern Europe, this seminar also offered the opportunity to gain information on the language resources of languages which are envolving to become more

relevant in the growing and more and more differentiating language technology market.

The collection of articles from the TELRI Seminar illustrates that there is a wide range of activities and topics in the domain of language resources and language technology. Analogue to the structure of the seminar different types of contributions are to be found here which cover the following different aspects:

- Information on trends and developments concerning fundamental topics, e.g. terminology, lexical resources, machine translation, public domain generic tools, and speech recognition.
- An overview on the developments concerning language technology and language resources in special locations (Europe and China).
- Information on international standardisation activities for corpora (the "Text Encoding Initiative").
- Information on activities concerning multilingual text tools and corpora for Central and Eastern European Languages (the "MULTEXT EAST" project).
- Information on activities (Concerted Actions and Joint Research projects) which are supported by the European Commission.
- Presentation of concrete (joint-venture) projects.
- Description of computer demonstrations given at the seminar.

The seminar was the first of its kind organised by TELRI, and we think that we can count it as a success. The presentations and demonstrations provided a broad fundament for discussions in the plenary sesssions as well as for direct communication and contacts. Two other TELRI Seminars are planned within the next two years and we think that this first one was a good beginning.

We hope that this presentation will be a useful source of information and, more over, that it will encourage engagement in co-operation activities and stimulate new ideas for projects and products in the field of language processing.

I would like to thank Lucie Piro, Joyce Thompson, and Norbert Volz for their logistic and editorial assistance.

Heike Rettig

Cooperation with Central and Eastern Europe in Language Engineering

Poul Andersen

DG XIII/E/6
European Commission
L-2920 LUXEMBOURG
Tel.: +352 4301 34324
Fax: +352 4301 34655
E-mail : poul.andersen@eurokom.ie
or Poul.Andersen@lux.dg13.cec.be

1. Introduction

The right of all citizens to communicate and receive information in their own language is a basic principle of the European Union. In order to preserve a multi-lingual Europe, with at present 11 official languages in the 15 Member States of the European Union, the European Commission feels naturally called upon to take a strong interest in Language Engineering (LE) along at least two dimensions:

- As a multilingual institution, it is itself an important user of LE products: The European Commission's Translation Service and other services use and support the development of Machine Translation and Machine-Aided Translation, Terminological Data Banks, and other tools that can help the Commission, especially in the huge task of translating more than 1 million pages each year.
  The Commission's own use of LE products is, however, outside the scope of this presentation.

- The European Commission encourages and financially supports research and development in the area of Language Engineering, with respect to the principle of subsidiarity, which implies that support mainly is given to multilingual activities, involving cooperation between partners from several member states.

2. Fourth Framework Programme

At present, scientific and technological cooperation in the field of LE is supported through the *Fourth Framework Programme for Research and Technological Development (1994–1998)* (4th FWP). This programme comprises a wide range of areas with a total budget of approximately 12 300 million ECU. The largest part of the funding is allocated to *Activity 1*, which covers RTD (Research and Technological Development) and Demonstration Programmes within the European Union. One of the Specific Programmes within Activity 1 of 4th FWP is the *Telematics Applications Programme*, which includes *Language Engineering* with a budget of approximately 81 million ECU.

The aim of Language Engineering is to facilitate the use of telematics applications and to increase the possibilities for communication in and between European languages by integrating new spoken and written language-

processing methods. Work focuses on pilot projects that integrate language technologies into information and communications systems and services. A key objective is to improve their ease of use and functionality and broaden their scope across different languages.

Language Engineering as defined in 4th FWP covers the following *Action Lines*:

1. **Pilot Applications**
   Document Creation and Management
   Information and Communication Services
   Translation and Foreign Language Acquisition
2. **Re-usable Language Resources**
3. **Language Engineering Research**
4. **Support Issues Specific to Language Engineering**
   (i.e. standards, assessment and evaluation,
   awareness activities, user surveys)

Since 1995, the Telematics Applications Programme, including Language Engineering, has been open to participation of institutions from Central and Eastern Europe, who can receive funding from the budget allocated to Activity 2 (see below).

The Third Call for Proposals under this programme was published on September 15, 1995, with a closing date on January 15, 1996. A final call is planned for publication on September 15, 1996.

A home page for Language Engineering can be found at:

HTTP://www.echo.lu/programmes/en/LangEng/le.html

All specific inquiries regarding Language Engineering within 4th FWP can be obtained from:

European Commission
DG XIII-E-5 LE Office
Batiment Jean Monnet (B4-002)
L-2920 Luxembourg
Fax: +352 4301 34999

2.1. International cooperation in the Fourth Framework Programme

A part of the Framework Programme of special interest to institutions in countries outside the European Union, is Activity 2, which covers Cooperation with *third countries and international organisations* with a total budget of 540 million ECU. Almost half of this budget is used to support cooperation with *Central and Eastern Europe.*

BEST COPY AVAILABLE

The main goals for RTD cooperation with Central and Eastern Europe are:
- to help to safeguard the RTD potential in these countries;
- to help to solve important social, economic, and ecological problems;
- to intensify cooperation in RTD fields where these countries are in the forefront on a world level.

The precise list of countries from Central and Eastern Europe that are eligible for support may change with the political development, e.g., in ex-Yugoslavia. In the latest two calls for proposals 1994 and 1995/96, the following groups of countries could participate:
- Countries of Central Europe (CCE):
  Estonia, Latvia, Lithuania, Poland, Czech Republic, Slovakia, Hungary, Slovenia, Romania, Bulgaria, Albania.
- Newly Independent States (NIS):
  Russia, Belarus, Ukraine, Moldova, Armenia, Georgia, Azerbaidjan, Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan, Uzbekistan.

The normal procedure for allocating funds is through *Open Calls for Proposals*.

A Call for Proposals was published on *October 17, 1995* with a deadline for receipt of proposals on *February 29, 1996*. A second Call for Proposals is planned for publication on *March 15, 1997*. These Calls are published in the EC Official Journal.

In order to participate in a Call for Proposals, it is necessary not only to read the text of the call in the Official Journal, but also to obtain the *information package* for the call, which contains detailed and specific information about what kind of proposals are eligible, as well as application forms to fill in with scientific, financial, and other administrative information. Such information packages are sent out to potential proposers, already known to the European Commission, and to anyone who asks for them.

The information package for the latest call with closing date on September 29, 1996 invites proposals for projects within Language Engineering, which ...
... must contribute to an open language infrastructure between the EU and CEC/NIS by focusing on:
- *creation of new language resources for the CEC/NIS languages;*
- *augmentation and further development of existing language resources for these languages;*
- *validation and exploitation of such resources for later integration into a range of computer-based services and products such as document management and translation tools which specifically identified needs in CEC/NIS.*

It also invites proposals for *Support Actions*, such as Awareness Seminars (see below).

Information packages are not always available in printed form at the date of publication of a call, but it is possible to collect this information over the INTERNET from

http://WWW.cordis.lu/

which contains an *Electronic Document Delivery Service* for documents and other texts related to the Fourth Framework Programme.

3. Implementation

Most of the funds under 4th FWP are spent on 3 types of activities:
- Concerted Actions
- Joint Research Projects
- Accompanying measures

3.1 Concerted Actions

In order to promote the creation of networks of scientists in the public and private sectors, the Commission supports *Concerted Actions*, which bring together teams from Eastern and Western Europe. Such Concerted Actions make it possible to establish permanent cooperation links which can serve as a basis for all kinds of research activity, and they encourage interactions between various disciplines, transfer of technologies, dissemination of results, and exchange of information in general. They encourage cooperation between academies and industries, help to identify new partners, and put research workers in contact with each other and with the responsible authorities in the different countries.

The intervention of the Commission covers coordination expenses: meetings, workshops, distribution of information, and exchange with and visits to other institutions taking part in the action. Financing can also be given for centralised facilities such as data banks, specialised communication facilities, and preparation and distribution of reference materials. They normally do not include funding of research, which the participating institutions are expected to get funded by other means, e.g., through EU-funded Joint Research Projects (see below).

By their nature, Concerted Actions are typically 'flat' structures with many participants from different countries.

The last Call for Proposals ***(COPERNICUS 1994)*** was published on January 31, 1994 with a deadline for receipt of proposals on April 29, 1994, and resulted in the funding of two Concerted Actions involving Central and Eastern Europe in the area of Language Engineering, both with Romanian participation:

- **TELRI – Trans-European Language Resources Infrastructure**
  TELRI has participants from all 11 *Countries of Central Europe*, as listed above. TELRI is coordinated by the *Institut für deutsche Sprache*, Mannheim, Germany, represented by Dr. Wolfgang Teubert. The other Western European participants represent leading institutions in Great Britain, France, Italy, the Netherlands, and Sweden, most of which are involved in networking activities within Western Europe such as ELRA *(European Language Resources Association)*, and can thus serve as a link between such activities and Central European institutions.
- **ELSNET goes East**
  ELSNET goes East is an extension of ELSNET *(European Network in Language and Speech)* to Central and Eastern Europe. Its geographical coverage differs from TELRI, as it does not have participants from all 11 Central European countries; however, it has participants from Russia and Belarus.

3.2 Joint Research Projects

These projects aim to assemble, for a specific research subject, a multinational team to perform research and development work and to obtain results in collaboration. A Joint Research Project typically comprises 3–6 partners coming from at least 3 different countries in Eastern and Western Europe. The duration of a project is normally between one and three years.

The above mentioned Call for Proposals, *COPERNICUS 1994*, resulted in the funding of the following 10 Joint Research Projects involving Central and Eastern Europe in the area of Language Engineering.

3.2.1. 'Broad' projects with many partners
→ good potential for creating infrastructure and networking in specific areas

Two Projects Within Speech:

- **ONOMASTICA-COPERNICUS**
  ONOMASTICA builds a pronunciation lexicon for the European Union, with city and town names, street names, family names, product names in 11 languages - Danish, Dutch, English, French, German, Greek, Italian, Norwegian, Portuguese, Spanish, and Swedish.
  ONOMASTICA-COPERNICUS extends the languages covered in ONOMASTICA to include names and pronunciations for *Czech, Estonian, Latvian, Polish, Romanian, Slovakian, Slovenian, and Ukrainian.*
  Pronunciation dictionaries for up to 250 000 names per language will be constructed, and quality controlled pronunciation lexicons will be made available in machine readable form (CD-ROM) for use in automated language systems by international European companies in the telecommunications sector and in the European (dictionary) publishing industry, as well as by language system researchers and developers.

- **BABEL - A Multi-Language Database**
  BABEL pursues the creation of large speech data collections for *Bulgarian, Estonian, Hungarian, Polish and Romanian:*
  - a speech database containing both read (about 75%) and spontaneous (about 15%) utterances collected from 100 speakers;
  - some spontaneous utterances will be accompanied by subsequent read versions (about 10%). It is planned to produce one CD-ROM disk per language (around 6 hours of speech).
  - a text corpus containing the orthographic text of the read utterances for each speaker.

  The project aims to produce phonetically and prosodically labelled annotations of at least 15% of the recorded material, using as far as possible semi-automatic labelling techniques, but ensuring expert checking.

One Project within Corpus Linguistics:

- **MULTEXT-EAST**
  MULTEXT-EAST is a spin-off of one of the largest EU projects in the domain of language tools and resources, MULTEXT, which had three main objectives:

BEST COPY AVAILABLE

- *Standardization*: development of a software standard based on a "software Lego" approach for corpus handling tools, together with TEI-based encoding conventions specifically suited to multilingual corpora and language engineering applications.
- *Tool and corpus development*: development of an extensive set of tools for corpus annotation and exploitation as well as the first annotated large-scale multilingual corpus for EU languages, intended to serve as a reference and test-bed for multilingual tools and applications.
- *Industrial validation*: integration by six major European companies of project results into high-level NLP applications such as term extraction and machine translation lexicon generation, thus providing a first indication of downstream applicability.

MULTEXT-EAST extends MULTEXT by transferring its expertise, methodologies, and tools to CEE countries, and the two projects together create a network of more than twenty academic research centers and companies, developing and using common lingware and methodologies, as well as producing the first annotated large-scale multilingual corpus for 12 EU and CEE languages.

East European languages covered: Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovenian.

One project within Terminology:

- **PRACTEAST - Preparatory Actions for Terminological Assistance to Central and Eastern European Countries**.

The concrete outcome of PRACTEAST is the compilation of four terminological collections in the domains of Economics and Management, Energy, Environment, and Telecommunications, each collection containing the 2000 most common terms with English terms and definitions and French and Spanish equivalent terms.

The databases compiled by the Coordinating Contractor, will revert to the CEC or to the infrastructure created as a result of the tasks completed within the framework of other EU funded projects. For example, the EURODICAUTOM database could be enriched with no less than 11 new European languages.

Furthermore, each partner will be able to use the Multilingual Database for its own research purposes and to publish the corresponding conventional paper dictionaries.

East European languages covered: Bulgarian, Czech, Estonian, Hungarian, Latvian, Lithuanian, Polish, Romanian, Russian, Slovakian, and Ukrainian.

3.2.2 'Narrow' projects with few partners
→ good potential for creating resources and/or applications in specific areas

Three projects concerned with Dictionary standards + coding:

- **CEGLEX - Central European GeneLEX model**
  CEGLEX aims at extending the generic electronic dictionary model (and accompanying SGML DTD) developed in the EUREKA GENELEX project to three central European languages and to:
  - give rapid access to a Western European linguistic engineering pre-standard model for Central European actors of the NLP scene,
  - extend the GENELEX model to new languages, evaluating its appropriateness and making it a stronger candidate for being an internationally recognised standard,
  - start a larger cooperation between the Partners leading to industrial level applications.

  The work will consist in
  - identifying theoretical issues in *Czech, Hungarian, and Polish* that may lead to specific adaptations and extensions of the model;
  - building a representative core lexicon that will conform to the elaborated model;
  - verifying the possibility to use the core dictionary in the context of an application.

- **GRAMLEX**
  The aim of GRAMLEX is to facilitate the initiation, coordination, and standardisation of the construction of morphological dictionary packages for French, *Hungarian,* Italian, and *Polish,* including detailed formal description of the morphology of the languages. The major challenges in such an enterprise are to give the description the largest possible coverage in order to be able to process unrestricted text; to share as many as possible of the formats, methods, and algorithms; and to improve time and space efficiency of programs.
  Keywords: lexical tagging, morphological dictionaries, lexical resources.

- **BILEDITA - Bilingual electronic dictionaries and intelligent text alignment**
  The goals of BILEDITA are
  - to provide a uniform dictionary format for all of the existing dictionaries constructed by the partners;
  - to provide a uniform lexical encoding scheme for both form and con-

tent of the entries in the electronic dictionaries. This has been systematically dealt with as far as the French and German dictionaries are concerned, and will be accomplished for the other project languages: *Bulgarian, Polish, and Russian;*
- to systematically construct and exploit bilingual corpora for the purpose of building bilingual basic dictionaries, terminology dictionaries, and phrasal dictionaries.

Two projects within CALL (Computer Assisted Language Learning):

- **BALTIC – Basic and Advanced Language Transnational Interactive Course**
  The objective is to create modular courseware for computer-assisted teaching of English to the citizens of *Latvia, Estonia, and Lithuania,* that will allow self-teaching, classroom teaching, and long distance teaching via an on-line network. BALTIC uses an already existing Italian/English course as a base, implementing the parts that allow the passage to other languages.

- **GLOSSER**
  GLOSSER applies NLP techniques, especially morphological processing and corpora analysis, to technology for computer-assisted language learning (CALL) with potential spin-offs in translation technology, information retrieval, and text-indexing technologies.
  GLOSSER's aim is to enable speakers of *Bulgarian, Hungarian, or Estonian,* who are intermediate language learners/users of English, to read and learn English more fluently. For example, when he reads a software manual on the screen and encounters an unknown word or an unfamiliar use of known word, he can point to it with the mouse and invoke online help, which will provide him with the following facilities:
  - a morphological parse, separating stem and ending, together with an explanation of the significance of the inflection;
  - the entry to the word in a bilingual X/English or a monolingual English dictionary;
  - (for a small number of words) an audible pronunciation;
  - access to similar examples of the word in online bilingual corpora.

One Speech project:

- **SQEL – Spoken Queries in European Languages**
  SQEL aims at the development of a multi-lingual and multi-functional information retrieval system, based on an existing, experimental infor-

mation dialogue system for English, German, French, and Spanish, SUNDIAL. Within SQEL, a prototype of such a system will be developed for *Czech, Slovak, and Slovenian.*

3.3 More Information on Concerted Actions and Joint Research Projects

A common home page for all 10 Joint Research Projects and the 2 Concerted Actions from the 1994 Call for Proposals can be found on the INTERNET, under

http://www.fwi.uva.nl/research/illc/ege/cop.le.proj.html

3.4 Accompanying Measures and Support Actions

Preparatory, accompanying, and support measures comprise i.a. the following activities:

- **National or Regional Awareness Seminars** are being conducted in Riga (for the three Baltic states), Prague, Poznan, Bucharest, and St. Petersburg in 1994-1996, and more are planned for other Central and Eastern European countries and regions. These seminars are aimed at opinion formers, media, providers, the research/academic community, users, and government organisations. Approaching these groups to make them realise the benefits of undertaking initiatives in the language engineering field, and the risks of not doing it, are some of the key subjects.
- **Information gathering** is supported, in order to identify possible cooperation partners in the area of language engineering in Central and Eastern Europe, and subsequently information dissemination to these partners. This information gathering is an important side-effect of the Awareness Seminars, and the Commission's recent, more systematic strategy towards the integration of Central and Eastern Europe into Trans-European cooperation activities started with a small seminar, which took place in Luxembourg in January 1994 and resulted in a first overview over the state of affairs in most of the countries.
  The seminar in Luxembourg was also used for the launch of a study carried out by ELSNET for the European Commission; the results of which were published in October 1994 as *Survey of Language Engineering Organisations in Central and Eastern Europe*. This document contains profiles of over 100 language engineering organisations in the following Central and Eastern

Europe and Newly Independent States: Belarus, Bulgaria, Czech Republic, Estonia, Georgia, Hungary, Latvia,Lithuania, Poland, Romania, Russia, Slovakia, Slovenia, Ukraine.

The document is available over the INTERNET at

http://www.cogsci.ed.ac.uk/elsnet/survey/survey.html

This survey is obviously not complete, but the information gathering is continued through the abovementioned Concerted Actions TELRI and ELSNET goes East, and will eventually result in an updated and more detailed picture.

4. More Information ...

about Calls for Proposals under Activity 2 of 4th FWP can be found

- on the Web at *http://www.cordis.lu* (as mentioned above);
- from local contact persons in each of the 11 eligible Central European countries and in Belarus, Russian Federation, Ukraine, and Georgia;
- directly from the European Commission:

Grazyna WOJCIESZKO
DG XIII
Tel.: +32-2-295 83 57
Fax: +32-2-296 17 16
E-mail : gwoj@dg13.cec.be

Inquiries about Trans-European cooperation activities related to language engineering may also be addressed to the author of this presentation.

An electronic mailing list *Eastern (Europe) Language Engineering* comprises more than 100 persons from Western as well as Central and Eastern Europe with a special interest in Trans-European cooperation. This mailing list can be used for announcement of conferences and other events, and any member of the list can enter search for cooperation partners in specific projects etc.

In order subscribe to this list, please send an E-mail to

poul.andersen@eurokom.ie

Language Technology and Language Resources in China

Feng Zhiwei

Institute of Applied Linguistics
State Language Commission of China
Chaoyangmen Nanxiaojie 51,
100010 Beijing, China
Fax: +86 106 513 8634

We are advancing into a new epoch – the information epoch. The remarkable feature of this information epoch is that the playing role of the computer in every aspect of society becomes more and more great. The natural language is the most important tool for communication of people, so it links an indissoluble bond with the information processing. In the information epoch, the computer with only a fourty year history gives a challenge to the Chinese language with six a thousand years history.

The Chinese language is the most important language of the Sino-Tibetan language family. Now nine hundred fourty million people in the world take Chinese language as their mother tongue. Not only Chinese people speak the Chinese language, some peoples in Singapore and Malaysia also speak it. The Chinese language is one of the working languages for the United Nations.

The Chinese character is the symbol set for recording the Chinese language. It is the largest symbol set of any writing system in the world. The Latin alphabet includes only 26 symbols, the Slavic alphabet 33 symblos, the Armenian alphabet 38 symbols, the Tamil alphabet 36 symbols, the Burmese alphabet 52 symbols, the Thai alphabet 44 symbols, the Lao alphabet 27 symbols, the Tibetan alphabet 35 symbols, the Korean alphabet 24 symbols, the Japanese Kana alphabet 48 symbols. However, there are too many symbols included in the Chinese characters. In the development procedure of Chinese characters, from ancient times to present times, the number of Chinese characters increased more and more. Following is the number of Chinese characters included in the dictionaries during different times of the Chinese history:

time	numbers
100 A.D. (Han Dynasty)	9353
543 A.D. (South Dynasty)	16917
1008 A.D. (Song Dynasty)	26194
1615 A.D. (Ming Dynasty)	33179
1716 A.D. (Qin Dynasty)	47043
1914	48000
1971	49888
1990	54678

With this big character set, how can the Chinese Natural language be processed by the computer? It is a great challenge to computational linguistics and corpus linguistics.

Forty years ago, in 1956, a Chinese scholar Ding Xilin suggested the creation of Šan, an electronic typewriter of Chinese characters. Prof. Qian Wenhao published a paper "Chinese characters and communication" in <Bulletin of Sciences>, in which he discussed the problem of encoding Chinese characters. In 1959, some Chinese scholars designed a machine translation system from Russian to Chinese (RC-59). It is the first connection between the Chinese language and the computer. In 1974, a large-scaled project, the 748 Project, was started: Chinese language processing became a new scientific subject in China.

In this paper, we shall introduce some results of Chinese language processing in China.

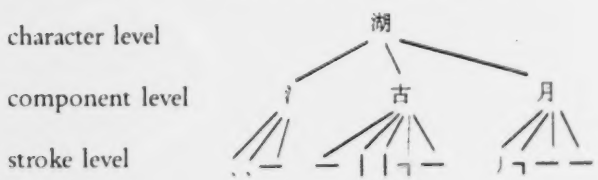
- Input of Chinese characters and Chinese corpus
- Automatic Segmentation of Chinese written text in corpus
- Grammar Knowledge Base for Chinese Words (GKBCW)
- Automatic POS (Part of Speech) tagging for Chinese corpus
- Automatic phrase bracketing and syntactic annotation for Chinese corpus
- Terminology Data-Banks
- Machine translation systems

1. Input of Chinese characters and Chinese corpus

The structure of a Chinese character can be divided in three levels:

character level
↑
component level
↑
stroke level

For example, 湖 (lake) can be divided as follows:



BEST COPY AVAILABLE

The Chinese character 潮 includes three components "氵古 月", and every component can include several strokes. The character is the higher level, the component is the middle level, the stroke is the lower level.

In the higher character level, the number of symbols is numerous, because different Chinese characters must be represented by different symbols. In the lower stroke level, the number of symbols is few, because we can sum up all strokes to several types. We can use only 5 kinds of strokes to represent all Chinese characters: horizontal stroke (–), vertical stroke (|), left-falling stroke (丿), dot stroke (丶), turning stroke (乛).

In the middle component level, the number of symbols is not so great, however, it is also not so few. The statistical result shows: 11 834 different Chinese characters are composed only by 648 different components, 16 339 different Chinese characters are composed only by 673 different components. If we further sum up these different components to several component categories and encode all Chinese characters by fewer component categories, the component categories can be distributed in the keyboard of typewriter: we can have the possibility to input Chinese characters by the universal keyboard of the typewriter (QWERTY keyboard).

Of course, we can also encode all Chinese characters by 5 strokes, thus we can also input Chinese characters by the universal typewriter keyboard. In this case, we need only 5 keys on the keyboard.

How do you input Chinese characters into a computer? Now we have a simple method – the encoding method. We can encode Chinese characters by their strokes or components.

We can also encode the Chinese characters with the aid of the phonetic alphabet. In 1958, the Chinese government published "The Phonetic Scheme for Spelling Chinese Language". In 1982, this Scheme was accepted by ISO/TC46 as an International Standard – ISO 7098. We call this scheme PINYIN.

If we take PINYIN to represent Chinese characters, we can directly use the typewriter keyboard very well, because there is very good correspondence between PINYIN and the universal keyboard.

The problem for PINYIN input is homophone. The number of total syllables in the Chinese language is 408. These 408 syllables represent a large number of Chinese characters, so homophone is inevitable. However, many Chinese words are polysyllable words (bisyllable or trisyllable). If we take the word as an input unit, the number of homophones will be obviously decreased. Now a great many Chinese prefer the PINYIN input. I believe that PINYIN input will become the main tendency of keyboard input method for Chinese characters.

For information interchange, the State Bureau for Technique Supervision

published the national standard "Character Set of Chinese Characters Encoding for Information Interchange - Basic Set" (GB 2312-80) in 1981. This standard includes two level of Chinese characters: the first level (commonly-used characters) includes 3755 characters; the second level (quasi-commonly-used characters) includes 3008 characters; the total number of characters included in GB 2312-80 is 6763. This is the basic set of Chinese characters.

In addition to the keyboard input method, there are the OCR input method and the speech-recognition input method.

OCR input method - The Optical Character Recognition (OCR) transforms the informations of Chinese characters in the paper into discrete electronic signals, then the computer recognizes these discrete electronic signals in order to recognize Chinese characters and to input Chinese characters into the computer. The Chinese OCR system can recognize 6763 Chinese characters in GB 2312-80; the recognition rate can reach 99.65%. The recognition speed can reach to 20 characters per second (in 386 PC).

Speech-recognition input method

- The computer directly transforms Chinese speech into Chinese text. The "SIDA-863A" system can recognize 398 basic Chinese syllables. The recognition rate can reach 93%. The response time is less than 0.1 seconds; input speed can reach up to 80 Chinese characters per minute. The syllable number of Chinese (420 syllables without ton, 1300 syllables) is fewer than English (4030 syllables) and Russian (2960 syllables), so Chinese speech recognition is relatively easy.

Since 1979, numerous Chinese corpora were created in China.

- corpus on Chinese contemparary literature (1979), 5.27 million Chinese characters, Wuhan University.
- Comprehensive Chinese corpus (1983), 20 million Chinese characters, Beijing Aviation & Spaceflight University.
- Corpus on Chinese language teaching materials for middle school (1983), 1068 million Chinese characters, Beijing Normal University.
- Comprehensive Chinese corpus (1983), 1.8 million Chinese characters, Beijing Language & Culture University.
- Corpus on Chinese Newspapers (1988), 2.5 million Chinese characters, Shanxi University:
  <People's Daily>: 1.5 million Chinese characters
  <Beijing Newspaper on Science>: 200 thousand Chinese characters
  <TV News> (CCTV): 500 thousand Chinese characters
  <Present Age> (magazine): 300 thousand Chinese characters.

- Chinese Corpus of Beijing University, begun in the early part of 1992, 5 million words.
- Bilingual (Chinese-English) corpus on computer science, begun in 1995 and is developing, State Language Commission, Applied Linguistics Institute.

From December 1991, a Chinese national Corpus project (sponsored by the State Language Commission) began to be put into effect. The purpose of this project is to build a large scale general corpus used for research of Chinese morphology, syntax, semantics, and pragmatics. This Corpus plans to collect 70 million Chinese characters.

The selection of this Corpus has the following restrictions:
- Diachronic restriction: to select the full range of materials from 1919 until now, and give priority to the material after 1977.
- Cultural restriction: mainly to select material that can be understood by the persons who have received formal education to the graduation of secondary school.
- Usage restriction: mainly to select the commonly-used material, and give priority to the material of social sciences and humane studies.

At present, this project progresses smoothly. By the end of 1995, the completed corpus will reach 20 million Chinese characters.

2. Automatic segmentation of Chinese writting text in corpus

A Chinese sentence is a sequence of Chinese characters, there are no obvious delimiting markers (such as spaces in Europaen languages) between Chinese words except for some punctuation marks. Because of this, word segmentation is essential in Chinese language processing.

In order to identify the embedded words in Chinese text, we match the input Chinese character string with the lexical entries in a large Chinese dictionary.

There are many matching methods:
- Maximum Matching method (MM method): taking 6-8 Chinese character string as maximum string, to match the maximum string against the lexical entries in dictionary, if failed, to cut one Chinese character and to match further until a corresponding word in the dictionary is found. The segmentation direction is from right to left.
- Reverse maximum matching method (RMM method): segmentation direction is from left to right. The experiment shows: RMM method is better than MM method.

- Bidirection Matching method (BM method): Comparing the segmentation resulted of MM and RMM, then to decide a correct segmentation.
- Optimum Matching method (OM method): In dictionary, the entries order was arranged by their frequency in the Chinese text, form higher frequency word to lower frequency word.
- Association-Backtracking method (AB method): By means of an association mechanism and backtracking mechanism to do the matching.

However, some ambiguous segmentation strings (ASSs) and unregistered words (the words that are not registered in the dictionary, URWs) in the text shall lower the accuracy of segmentation.

There are two types of ASSs:
- overlapping string: e.g., 太平淡 (very prosaic, rather flat): 太平 (peace and tranquility) + 平淡 (prosaic), 平 becomes overlapping segment.
- combinative string: 马上 e.g., (at once): 马 (horse) + 上 (upper) = on the horse.

URWs are mainly proper nouns: personal name and place name. e.g, 冯志伟 (Feng Zhiwei), 赫豪尼 (Tihany). These are not included in the dictionary.

To resolve these problems, various knowledge might have to be consulted. Knowledge on the part of speech (POS) will be helpful. If we blend the POS annotation and automatic segmentation, the accuracy of segmentaion will be increased remarkably.

For Chinese word segmentation, a natioanl standard was promulgated in 1992 - 'Contemporary Chinese language word segmentation standard used for information processing' (GB 13715). This national standard proposes the principle for determination of Chinese words. It is the basic principle for automatic word segmentation.

3. Grammar Knowledge Base for Chinese Words (GKBCW):

Now a GKBCW is developing in Peking University. This GKBCW can be used as a large electronic dictionary; it is a support to the automatic segmentation and Chinese Corpus annotation.

In GKBWC, each word category has several features, e.g., the verb category has about forty features, the noun category has twenty-five features. These features describe the grammatical functions and distribution of every Chinese word in the text. GBKCW includes about 50 000 word entries which composed a general base and 27 sub-bases.

The general base describes the common features for all words, such as
- pronunciation of word: e.g., 太平 [tai4ping2], 马上 [ma3shang4].
- part of speech of word: e.g., 编辑 can be verb (to edit) or noun (editor), 制服 can be verb (to bring under control) or noun (uniform).
- type of ASSs: e.g., 太平 can lead to overlapping string, 马上 can lead to combinative string.
- usage features: e.g., frequency, subject domain, style, rhetoric feature.
- special information: e.g., radicals of Chinese character by which characters are arranged in traditional Chinese dictionaries. 纟 in Chinese character 编 is a radical in the traditional dictionary.

The 27 sub-bases are for basic parts of speech, idiom, some Chinese characters which can not be used as morpheme, and Chinese punctuation markers, etc.

The general base is linked with 27 sub-bases. All the information in the general base can be transfer to each sub-bases.

4. Automatic POS tagging for the Chinese Corpus:

There are two kinds of POS tagging approach:
- Statistics-based approach
- Rule-based approach

The processing procedure for statistics-based approach of POS tagging can be divided into the following steps:
- Manually analyse some texts selected from corpus (training set), annotate the training set, and extract the statistical data from the analysed training set (it is represented by 2-tuple grammar);
- construct a statistical model according to the results from the statistical data of training set;
- automatically annotate new texts based on the statistical model.

All tagging information is coming from the electronic dictionary where is recorded the information about the POS for every word.

The automatic POS tagging system of Qinhua Unversity in Beijing takes the statistics-based approach; the accuracy of POS tagging reached 96.8%; the annotating speed is 175 occurances per second.

For the rule-based approach, a serious problem is the POS ambiguity. In Chinese language, POS ambiguity mainly concentrates on the frequently-used words: verb, noun adjective, etc.

verb-noun ambiguity:	37.60%
verb-adjective ambiguity:	24.30%
noun-adjective ambiguity:	10.40%
adjective-adverb ambiguity:	4.55%
verb-preposition ambiguity:	4.04%
verb-adverb ambiguity:	2.27%
noun-verb-adjective ambiguity:	2.27%
noun-adverb ambiguity:	2.02%
other ambiguity:	12.55%

The disambiguation must be based on linguistic rules (grammar, semantics, context, etc), so that the information included in GBKCW will be very helpful.

In fact, the statistics-based approach is an empirical approach, and the rule-based approach is a rational approach. We can combine both statistics-based approach and rule-based approach into one, and integrate different types of approaches in POS tagging. By this approach, experimental results of Peking Unversity are:

- segmentation accuracy: 97.68% (close test)
- POS tagging accuracy: 96.06% (close test), 95.72% (open test).

5. Automatic phrase bracketing and syntactic annotation for the Chinese Corpus

After word segmentation and POS tagging for Chinese corpus, we must manually proofread the results, and when we can confirm that the results are good, then we can start the automatic phrase bracketing and syntactic annotation.

The procedure is as follows:

- To predict the boundary locations of a phrase according to the information about words, their POS and other syntactic features, and determine which word is the left boundary of a phrase, which is the right boundary of a phrase, which word is the middle part of a phrase.

We can bracket a phrase as following:

[w w ... w w]

[w is an open bracket, w] is a closed bracket.

- To match the open brackets and its corresponding closed brackets based on the context information.

- To resolute the ambiguity of the phrase according to the disambiguation rules and statistical information.
- To generate the constituent structure tree for a sentence.

Recently a Chinese Corpus Multilevel Processing system (CCMP) is being developed at Peking University. This CCMP system includes two sub-systems and supplementary tools:
- Word segmentation and POS tagging sub-system
- Phrase bracketing and syntactically annotating sub-system
- supplementary tools such as query tools, sample tools, statistical tools, and corpus management interface.

The experimental results:
- the percentage of crossing brackets: 13.98%
- the percentage of error phrase tags: 8.65%

That means there are many problems to resolve for the annotation of Chinese corpus.

6. Terminology Data-banks

The terminology is crystallization of scientific knowledge in language, it is an important language resource.

In 1990, the sub-committee of computer-aided in terminology of China was set up. This sub-committee is attached to the State Language Commission (SLC) of China.

A series of national standards for a terminology data-bank are promulgated:
- General principles and methods for establishing terminology data bank, 1992, GB/T 13725-92.
- Magnetic tape exchange format for terminological/lexicographical records, 1992, GB/T 13725-92.
- Guideline for the development of terminology data banks, 1993.
- Guideline for the documentation for developing terminology data bank, 1993.
- Guideline for the evaluation of terminology data banks, 1994.

Many terminology data-banks are created:
- GLOT-C: data processing termonology, Chinese-English, 1988, the Academia Sinica collaborated with FhG of Germany.
- TAL: applied linguistics termonology, Chinese-English, 10 000 terms, the State Language Commission, 1990.

- COL: computational linguistics terminology, Chinese-English, 10 000 terms, the State Language Commission collaborated with Trier University of Germany, 1993. (in press by Langenscheidt Verlag).
- Terminology data-bank on computational linguistics: Chinese-English-German-Japanese, 12000 terms, Peking Unversity, 1994.
- Terminology data-bank on machinary: 250 000 terms, Chinese-English-French-German-Russian-Japanese, Institute for Scientific and Technical Information, the Ministry of Machinary, started since 1989, in development.
- Thesaurus bank on agriculture: Chinese-English, 25 000 terms, Chinese Academy for Agriculture, 1991.
- Thesaurus bank on chemical industry: Chinese-English, 25 000 terms, China Information Center of Chemical Industry, 1989. There are two versions: published version and machine-readable (floppy discs) version. All terms can be transmitted through the network to provide information retrieval service.
- Encyclopedia terminology data bank: Chinese-English, there are definition and explanation for every term, 180 000 terms, China Encyclopadia Press, 1995.
- Terminology data bank for standardization: Chinese-English, it includes several sub-databanks:
  . comprehensive terminology data bank (TDB)
  . comprehensive bibliographic data bank (BDB)
  . comprehensive factual data base (FDB)
  . filing system administration in the form of a comprehensive data-bank (FSA)
  . multiple documentation language system (MDL)
  . office automation system (OAS)
  . full text system (FTS)
  . data base for graphical and other non-linguistic data (GDB)
  China Standardization and Information-Classification-and-Coding Institute (CSICCI) collaborated with Oesterreichisches Normungsinstitut, in developing.
- Project of Comprehensive scientific terminology data bank: Chinese-English, 50000 terms, Institute of Scientific and Technical Information of China (ISTIC), the project started in 1995, shall fulfill in 1998.

7. Machine Translation Systems

The study of machine translation (MT) in China started over forty years ago. The development of MT in China can be described in four periods:
- the early experimental period (1956–1966)
- the stagnant period (1966–1975)
- the recovery period (1975–1987)
- the blossom period (since 1987)

In 1956, the MT research has been included in the National Plan for Developing the Science and Technology as a project named "Machine Translation – establishing of the translation rules of natural language and mathematical theory for natural languages". This project can be divided into two parts: one is the establishment of the translation rules of natural language – "machine translation", another is the study of mathematical theory for natural language – "mathematical linguistics" (the theoretical foundation for machine translation). Several research groups were founded in Beijing (Academia Sinica, Beijing Institute of Foreign Language), in Guangzhou (South China Polytechnical Institute), and in Harbin (Harbin Polytechnical University). In 1959, a Russian-Chinese MT experiment (RC-59) was successfully fulfiled on a general-purpose computer 104. With a vocabulary of 2030 Russian words, an algorithm of 29 flowcharts, this RC-59 experiment encouraged the belief that MT from foreign language into Chinese is possible. At the same time, some scientists began to study the MT from English to Chinese. In 1960, an English-Chinese MT algorithm was composed. A monographical brochura "Preliminary of Machine Translation" was published in 1965. However, from 1966 until 1975 the MT in China completely stagnated.

Since 1975, after a long sleep of 10 years, the MT in China restarted and came to the recovery period.

In November 1975, a MT joint-research group was established. This MT group consists of the Institute of Scientific and Technical Information of China (ISTIC), the Linguistics Institute of the Chinese Academy for Social Sciences (CASS), the Computational Technique Institute of Academia Sinica etc. This group carried out a MT experiment from English to Chinese on the TK-70 computer and T-4100 information processing device of Chinese characters. The raw materials contain 9200 English titles of scientific and technical papers. As the results of this MT experiment, a MT system TITLE-1 was set up in 1986.

At the same time, MT study was carried out in Helongjiang University (Harbin), in the Mars Institute (Beijing), in the Telecommunication

University (Beijing), in South China Polytechnical University (Guangzhou), in Central China Polytechnical University (Wuhan), in the Institute of Scientific and Technical Information of Shanghai (Shanghai), and in Inner Mongolia Unversity (Huhehot).

In this recovery period, interactive approaches and diverse strategies were developped, some multilingual systems appeared, and the application of AI technique in MT began to be considered. Mathematical linguistics were also studied in the universities or the institute. A monograph "Mathematical Linguistics" was published in 1985.

For the sake of investigation of linguistical phenomena, an English corpus was created in Jiaotong University (Shanghai), and numerous Chinese corpora were created in Wuhan University, Peking University, Qinhua University, Shanxi University.

In the recovery period, most of the MT system was experimental:

- TITLE-1 system: English-Chinese, ISTIC, 1976-1986.
- ECMT-1 sytem: English-Chinese, Linguistics Institute, CASS, 1978.
- JFY system: English-Chinese, Linguistics Institute, CASS, 1976-1984.
- INSPEC sysetm: English-Chinese, Telecommunication University, 1985.
- HT-83 system: English-Chinese, Helongjiang Unversity, 1983.
- RI-84 system: English-Chinese, Helongjiang University, 1984.
- GCAT system: German-Chinese, Applied Linguistics Institute, LSC, 1985.
- FCAT system: French-Chinese, Applied Linguistics Institute, LSC, 1985.
- FAJRA system: Chinese-French/English/Japanese/Rassian/German, Applied Linguistics Institute, LSC, 1981.

The TITLE-1 system possessed a large-scale electronic dictionoary including a basic dictionary (20 000 entries) and an idiomatic dictionary (67 000 entries). This system can translate the English titles of scientific papers in the field of metalurgy to Chinese, the average translation speed is 80 titles/hour.

Since 1986, the MT of China came to the blossom period. The symbol of this blossom period is the KEYI-1 English-Chinese system of the Mars Institute (Beijing). In March 1987, KEYI-1 system passed the academic appraisal by experts. Its translation ability is as the ability of graduated students of the English department in China, its translation speed is 3000 words/hour, and the result of translation is readable. In the process of machine translation, the user can input their special words to KEYI-1 in order to adapt to their special demands.

KEYI-1 system quickly became an operational system and was commercialized; China National Software & Technology Service Co. (CS&S) bought the copyright of the system, and KEYI-1 system was renamed as TRANS-STAR system. CS&S put it on the market and gained the profit.

Now TRANS-STAR system has been improved. It is now much better than KEYI-1. The translation speed is raised to 15 000 words/hour for 286 PC, 30 000 words/hour for 386 PC. The basis dictionary includes 40 000 entries; the system has 10 specialized technical dictionaries including 350 000 entries. The subject fields involved computer, economics, telecommunication, ceramics, thermal power industry, printing machinary, automobile and tractor industry, petroleum prospecting, geology, and chemical industry.

In the blossom period, another three operational systems are also very successful:

- GAOLI MT system (English-Chinese): It is jointly developed by Beijing GAOLI Computer Co. Lid. & Linguistics Institute of CASS.
- basic lexical base: 60 000 entries in which the usage of every word is described.
- linguistic rules: more than 800 rules used for syntactic analysis of English and generation of Chinese.
- background knowledge base: more than 150 entries used for semantic analysis and generation
- translation accuracy: 80%
- readability of translated text: 80%-90%
- translation speed: 12000 words/hour for 386 PC.
- 863-IMT/EC system (English-Chinese): It is developed by the Computer Technology Institute, Academia Sinica. This system was commercialized and earned very good economic benefits.

. basic English lexical base: 35 000 entries
. basic Chinese lexical base: 25 000 entries
. linguistical rules: 1500 rules
. translation accurancy: 80%

- SINO-TRANS system (Chinese-English): It is developed by CS&S at 1993.

. basic dictionary: 40 000 entries
. two specialized technical dictionaries: navel ships and boats (9312 entries), rocket gun (33 773 entries)
. linguistic rules: 1000 rules
. translation speed: 20 000 Chinese characters/hour.

Since 1989, the corpus approach (e.g., statistical approach, example-based approach) is introduced to machine translation, all the research work of machine translation are based on the processing of large-scale authentic corpus. The combination of machine translation with the corpus approach will promote the development of Chinese language technology. Corpus linguistics play more and more a role in Chinese language technology. The prospect of Chinese language technology will be more and more brilliant.

References

Feng Zhiwei. 1982. "Memoire pour une tentative de traduction multilingue du chinois en francais, anglais, japonais, russe et allemand" Proceedings for COLING'82, Prague.

Feng Zhiwei. 1983. "Multi-label and multi-branch tree for automatic analysis of Chinese sentences". Proceedings for 1983' International conference on Chinese information processing, Beijing.

Feng Zhiwei. 1984. "Automatic generation and analysis of Chinese language in machine translation". Proceedings of SEARCC'84, Hongkong.

Feng Zhiwei. 1987. "Linguistic information included in Chinese sentences". Proceedings of TKE'87, Trier.

Feng Zhiwei. 1989. "Some special problems of machine translation in China". Proceedings for Chinese Computing Conference'89, Singapore.

Feng Zhiwei. 1990. "Complex features in description of Chinese language". Proceedings for COLING'90, Helsinki.

Feng Zhiwei. 1990. "Automatic analysis of Chinese - MMT model". Proceedings of 1990 International Conference on Computer Processing of Chinese and Oriental language, Beijing.

Feng Zhiwei. 1991. "On potential ambiguity in Chinesec terminology". Proceedings of TSTT'91, Beijing.

Public Domain Generic Tools: An Overview

Tomaž Erjavec

Language & Speech Group
Intelligent Systems Dept.
Jožef Stefan Institute
Ljubljana, Slovenia

1. Introduction

This paper gives an introduction to language engineering software, especially as it relates to computerised textual corpora. The focus of the paper is on language engineering tools, i.e. relatively small and independent pieces of software, meant for a particular, usually low-level task. Other, larger and more complex systems will be mentioned as well, as long as they are connected to the processing of textual material, in particular to corpus production and, to some extent, its utilisation. The paper does not discuss tools dealing with speech production or recognition although some of the corpus tools are relevant for producing speech corpora as well.

Even though the focus of this article will be on public domain tools, some software will be mentioned that does not, strictly speaking, belong in this category. There is a substantial variety of conditions that authors impose on their software, with proprietary, commercial products on one end, and freely available public domain software, that can be used for any purpose at the other end of the spectrum. Quite a few interesting linguistic tools fall somewhere between these two extremes with the most common conditions being that the software may be freely used for non-profit purposes only or that it falls under the GNU's general public license, which essentially forbids such software from being incorporated into proprietary programs. Furthermore, some authors request an explicit license to be signed before releasing their software. Nevertheless, such systems, even though not in the public domain, can be used by academic users and, in certain cases, can be of substantial use in an industrial environment as well. So, for example, if the software publicly released for academic use only is of sufficient interest, an arrangement can usually be made with the authors, or, if the software is a GNU library, it can still be used by proprietary software, as long as it does so in accordance with the GNU library general public license.

Using public domain tools has several obvious benefits, which are probably greatest for smaller research teams. These often lack funds to buy proprietary software or manpower for in-house development. Besides the obvious benefit of being for free, public domain tools allow for exploring a particular technology; even if a tool is not exactly what is required, the source code (where available) can be modified to suit particular needs. With public domain linguistic software that incorporates language particular resources, these resources (e.g., a morphological rule-base) can also be reused locally. Of course, the problems associated with using public domain tools should not be underestimated. These, along with some future prospects for their resolution will be discussed in the concluding section. Finally, for many tasks, in particular

for corpus work, commercial software is simply not available. Therefore, the available options are narrowed to using (and possibly modifying) public domain tools, or re-inventing the wheel by in-house development of the software.

The rest of the article is organised as follows: first, some introductory remarks are given on corpora and their connection with standards and technological advances; next the Unix platform, as the preferred development environment is discussed; this is followed by a section on SGML and statistical tools and a section on computational linguistic tools. Finally, some drawbacks to using public domain tools are given, together with recent efforts in this filed, concluding with a list or Web sites relevant to language engineering tools.

2. Corpora, Standards and the Internet

Recent years have seen a steep growth of computer corpora which have increased in size, number, and variety. Two of the more impressive examples of this trend are the hundred million annotated words of the British National Corpus and the hundred CD-ROMs of language resources offered by the Linguistic Data Consortium, which include annotated spoken corpora and multilingual corpora, both from a variety of sources.

The increased ability to produce and disseminate corpora is to a large extent due to technological advances. The dropping price and large capacities of mass storage media mean that large and heavily annotated corpora can be easily stored on-line. The growth in electronic communications, especially the success of the World Wide Web enables information on language resources or the resources themselves to be offered and accessed globally. In addition, the growing acceptance of certain standards that enable the exchange and platform independence of corpora encodings have also had a important impact on corpora availability and reuse. In particular, the Text Encoding Initiative guidelines (Sperberg-McQueen&Burnard, 1994, Ide&Veronis, 1995), which adopt the ISO standard SGML (Goldfarb, 1990) as their markup (meta)language are a significant contribution to the standardisation effort in this area.

This ease of availability and adoption of standards is important not only for corpora themselves, but also for software that helps in producing and, to a lesser extent, utilising these corpora. Internet connections mean that such tools as are offered to the public can be easily down-loaded or, in some cases, demonstrated, while the increasing adoption of standards minimises portability and interface problems.

3. The Unix platform

Although publicly available software does exist for PCs and Macs, it is Unix that is practically based on the notion of free software. This is due as much to the developmental history of Unix as to the GNU initiative of the Free Software Foundation (FSF). Although GNU software is not public domain, it can be used and modified freely, as long as it is not incorporated into proprietary systems. The utility of Unix as a development environment is due to it being a very powerful system, with a "amorphous" structure, that imposes relatively few constraints on program developers. For these reasons, it will be primarily Unix software that will be discussed in this article.

It should be noted, however, that it is also because of its power and reliance on free software that Unix is in many ways a troublesome system to use and maintain. Furthermore, Unix comes in a thousand and one flavours, depending on the exact platform in use (e.g., Solaris for SUN Sparcs, Linux for PCs, Irix for SGIs, etc.), thus, often making the installation of new programs a difficult undertaking.

I list next some software which runs on Unix (but often on other platforms as well) and can be of use in language engineering. First, Unix offers a variety of tools that do a specific job well, for example string or regular expression searching (**grep**) or sorting (**sort**). If programming languages can be thought of as "tools", then Unix offers a large selection of use in writing programs for language processing. The general purpose programming language which is currently, and probably for a while to come, the de facto standard is ANSI C and it's object-oriented extension C++. Unix also offers a number of languages that are particularly suited for string processing, such as **sed** (Doucherty, 1991) and **awk** (Aho et al., 1988). **Perl** deserves particular mention (Wall&Schwartz, 1991): it is suitable as much for writing short, throw-away programs as for complex conversion tasks. Finally, the GNU editor, Emacs, must be mentioned which, although as with most things with Unix, has a long learning curve, does offer very powerful functionality, and is freely extensible in its variant of Lisp.

4. SGML and Statistical Tools

The view of the corpus building process adopted here revolves around (presumably TEI conformant) SGML as the underlying data representation format. The evolution of a corpus is seen as composed of three stages. The corpus texts will usually be obtained in some sort of machine readable legacy

format (e.g., an ASCII representation, RTF from Word files, etc.) which are first up-translated into a corpus-wide encoding format, i.e., SGML. This bibliographically and, to an extent, structurally encoded corpus is then usually additionally (SGML) annotated in a number of ways, for example, for part-of-speech or multilingual alignment. Finally, a corpus is utilised by *searching* and *rendering* its material, for example, by showing keywords in context that match a given criterion or by showing aligned multilingual texts side by side.

The Perl language is well suited for up-translation to SGML as well as most (non-linguistic) conversions of SGML documents; there also exists an SGML-aware Perl library perlSGML, written by Earl Hood. Another general purpose programming language which is in the public domain and particularly suited to the manipulation of character strings is Icon (Griswold&Griswald, 1990). It was developed at the University of Arizona, and was extensively used in the British National Corpus project.

The basic SGML tool is the validating parser that checks for syntactic well-formedness of SGML documents and reports errors in case the document is not well-formed. A number of such validators exist, quite a few of which are in the public domain, e.g., James Clark's **sgmls** and **sp**. The second essential "tool" that is needed is an SGML-aware editor. Most of these are commercial software; however, Emacs does have a special mode (**psgml**), meant for editing SGML files.

There also exist freely available programs specifically designed for conversion of SGML documents (although not designed for handling corpus data), for example, the Copenhagen SGML tool CoST and MID's **MetaMorphosis**. These allow for transformations of SGML documents and also for rendering SGML annotated data. These tools and others can be found in the various Internet SGML repositories.

While there are quite a few tools available for corpus development, the choice of corpus querying tools is much more limited. While some of the above tools might prove useful in designing such a system, an integrated corpus query system must combine speed, a powerful querying language, and a display engine. Some such systems have been developed for DOS, but they usually lacked support for non-English languages and relied on idiosyncratic corpus encoding schemes. One Unix system that is offered for research purposes is the Corpus Query System cqp/Xkwic by Stuttgart's Institute für Maschinelle Sprachverarbeitung (IMS). The corpus query processor **cqp** is a command-language based query interpreter, which can be used independently or by **Xkwic**, which is a X-windows graphical user interface.

The last part of this section mentions some statistical linguistic tools used for corpus annotation. Part-of-speech taggers take as their input a word-form

together with all its possible morphosyntactic interpretations and output its most likely interpretation, given the context in which the word-form appears. So, for example, the word-form "tags" by itself can be interpreted as the plural noun or as the third person singular verb, whereas in the context "it tags word-forms" only the verb interpretation is correct. While syntactic parsers also perform such disambiguation, pure rule-based approaches tend to have low coverage and speed, and the investment into building rulesets for a particular language is prohibitive.

Recently there has been an increased interest in statistically based part-of-speech taggers, which use the local context of a word form for morpho-syntactic disambiguation. Such taggers have the advantage of being fast and can be automatically trained on a pretagged corpus. Their success rate depends on many factors, but is usually at or below 96%. Two better known such taggers in the public domain are the Markov model-based Xerox tagger written in Lisp (Cutting et al., 1992) and Brill's rule-based tagger in C (Brill, 1992).

Finally, another pure statistical tool is the Gale and Church aligner, (Gale&Church, 1993) which sentence-aligns a text and its translation. It produces surprisingly good results by very simple means as it incorporates no linguistic knowledge but makes use of the basic insight that a text and its translation will have roughly the same number of characters.

5. Computational linguistic tools

This section deals with software that belongs to computational linguistics proper and includes morphological analysers, implementations of formalisms, and lexicon development environments. These systems can hardly be considered "tools" as they are often large and complex. They are, furthermore, only distantly connected to corpus development or exploitation. Nevertheless, they provide an environment for advanced language engineering tasks (e.g., machine translation) and it would be remiss not to mention them.

For morphological analysis and synthesis, Koskenniemi's finite-state *two-level model* is by far the most widely used and investigated. It is primarily meant to deal with spelling changes at or near morpheme boundaries. The best known implementation is probably PC-KIMMO (Antworth, 1990) although a number of other implementations also exist. Information about them, as well as about other systems, not based on the two-level model is available from the Saarbrücken's DFKI Natural Language Software Registry.

For general lexical structuring, including, but not limited to morphological dependencies, a simple, yet powerful and efficient language is DATR

(Evans&Gazdar, 1990). DATR is a lexical knowledge representation language in which it is possible to define networks allowing multiple default inheritance. The original Sussex version, which is publicly available, is implemented in Prolog.

Syntactic parsing, which usually forms the basis of more advanced language engineering applications has probably been the subject of most research in computational linguistics. It is, therefore, not surprising that a host of (public domain) parsing programs are available. For example, Prolog implementations usually offer a Definite Clause Grammar (**DCG**) module, and a number of various parsers (chart, Tomita, etc.) are available via the Internet, e.g., via **DFKI**.

Apart from **DCG**, the best known unification-based context-free parser is the **PATR** system (Shieber et al., 1983). More recent unification-based systems have replaced untyped feature structures of PATR with typed ones, thus, conferring the benefits of type checking and type inheritance to their grammars. Given that these systems can be used for other purposes apart from just parsing (e.g., machine translation), they are better classified as implementations of linguistic formalisms. There are a number of such systems available, pointers to which can be, again, found at the DFKI Web page. Here we will mention only three of the better known ones. The Attribute Logic Engine (ALE) (Carpenter&Penn, 1994) is written in Prolog and incorporates a chart parser and lexical rules. It is optimised for speed of processing which, however, makes it less than ideal for a grammar development environment. IMS offers two systems: Comprehensive Unification Grammar (CUF) (Dörre&Eisele, 1991) written in Prolog and Typed Feature Formalism (TFS) (Zajac, 1992) in Lisp. Especially CUF offers a very powerful grammar development environment; for a detailed comparison between ALE, CUF, and TFS, see also (Manandhar, 1993). Finally, it should be noted that of the three systems, ALE is available in source code, the other two being distributed in their compiled version only.

6. Drawbacks and Prospects

The penalties of using public domain tools should not be underestimated: the tools often do not come with all the bugs ironed out, with a detailed documentation or with the exact functionality required on the platform that we have. Maintenance is also often lacking as the developers' interests can have turned to other areas and support is, of course, a voluntary effort and cannot be counted on. For the field of multilingual language engineering, an

especially serious problem is that most of the available linguistic engineering software (public domain or commercial) to date was written for the English language or at best for (major) European Union (EU) languages. This bias gives rise to problems with "foreign" character sets, collating sequences, and the format of dates, numbers, and the like.

A connected problem is the lack of standards or, in some cases, conflicting standards for software development. This gives rise to tools that are often incompatible with one another, e.g., by virtue having different input/output formats and protocols. This can make their integration a daunting task, requiring extensive modifications of the tools. This is not to say that standards concerning software development have not been developed or are being considered by various institutions, e.g., by ISO and FSF. The work that is specifically addressed towards linguistic engineering are the *Guidelines for Linguistic Software Development*, which are being produced as a joint effort of the EU sponsored MULTEXT and MULTEXT-East projects and the Eagles sub-group on Tools, established in spring 1995. In particular, these guidelines are to address questions of usability, portability, compatibility and extensibility of linguistic software, concentrating in the first place on the Unix environment.

A number of other European Union projects have been concerned with developing linguistic software. However, in most cases the produced software is proprietary and, hence, not publicly available. A notable exception is the MULTEXT(-East) (this volume) project, which aims to make freely available to the academic community a number of SGML-based corpus processing tools. These include re-implementations of the already mentioned Xerox tagger, the Gale & Church aligner and a morphological synthesiser based on the two-level model.

For obtaining the tools mentioned in this paper, as well as a host of others, the easiest way is via the Web. A number of Web sites that provide further pointers to resources of interest to linguistic engineering, already exist: some of them are listed below. While some are the product of voluntary effort by individuals, there are also official bodies that disseminate information via the Internet, e.g., the DFKI Natural Language Software Registry or the EU Relator project. In connection with this, the pioneering effort of Edinburgh's Language Technology Group should also be mentioned: LTG offers a *Language Software Helpdesk*, which is a free service dedicated to the support of public domain, and freely available software for natural language processing and the fostering of its use in practical applications.

Finally, the TELRI Concerted Action also has a working group on "Lingware Dissemination". Its purpose is to increase the availability of language

engineering tools by making available, via the Web, information on extant tools, by providing the public tools of TELRI partners, tools, and by improving such tools by adapting them to various languages and platforms.

7. WWW References

The following World-Wide Web pages can provide more information on many of the topics and tools introduced above:

SGML Web Page by Robin Cover and SGML Open:
http://www.sil.org/sgml/sgml.html
http://www.sgmlopen.org/

TEI home page:
http://www-tei.uic.edu/orgs/tei/

British National Corpus and Linguistic Data Consortium:
http://info.ox.ac.uk/bnc/
http://www.cis.upenn.edu/ldc/

GNU software ftp site (including Emacs, Perl, grep, etc.) and online documentation for GNU software:
ftp://prep.ai.mit.edu/pub/gnu/
http://www.ns.utk.edu/gnu/

SGML repository at Institute for Informatics, Oslo (including psgml, sgmls and sp) and Steve Pepper's Whirlwind Guide to SGML Tools:
ftp://ftp.ifi.uio.no/pub/SGML/
http://www.falch.no/people/pepper/sgmltool/

Taggers by Xerox and Brill:
ftp://parcftp.xerox.com/pub/tagger/
ftp://blaze.cs.jhu.edu/pub/brill/Programs/

DFKI Natural Language Software Registry and the IMS list of language engineering links:
http://cl-www.dfki.uni-sb.de/cl/registry/draft.html
http://www.ims.uni-stuttgart.de/info/FTPServer.html

BEST COPY AVAILABLE

SIL's Software (including PC-PATR, PC-KIMMO):
http://www.sil.org/computing/sil_computing.html

DATR ftp site:
ftp://ftp.cogs.sussex.ac.uk/pub/nlp/DATR/

ALE home page:
http://macduff.andrew.cmu.edu/ale/

IDS's Tools and resources, including CQP/Xkwic, CUF and TFS:
http://www.ims.uni-stuttgart.de/Tools/ToolsAndResources.html

LTG's Language Software Helpdesk:
http://www.ltg.hcrc.ed.ac.uk/projects/helpdesk/

MULTEXT with Eagles Guidelines for Linguistic Software Development and MULTEXT-East:
http://www.lpl.univ-aix.fr/projects/multext/
http://www.lpl.univ-aix.fr/projects/multext-east/

Eagles and Relator:
http://www.ilc.pi.cnr.it/EAGLES/home.html
http://www.de.relator.research.ec.org:80/lg=en/index.mlhtml

TELRI and the Web version of this article:
http://www.ids-mannheim.de/telri/telri.html
http://nl.ijs.si/telri-wg5/pub-tools/

References

Aho, A.V., Kernighan, B.V., & Weinberger, P.J. 1988. "The Awk Programming Language". Addison Wesley, Reading, Massachussetts.

Antworth, E.L. 1990. "PC-KIMMO: a Two-level Processor for Morphological Analysis". No. 16 in Occasional Publications in Academic Computing. Summer Institute in Linguistics, Dallas, Texas.

Brill, E. 1992. "A Simple Rule-based Part of Speech Tagger". Proceedings of the Third Conference on Applied Natural Language Processing, ACL Trento, Italy.

Dörre, J. & Eisele, A. 1991. "A Comprehensive Unification-based Grammar Formalism". Technical Report Deliverable DYANA R3. 1.B, Centre for Cognitive Science, Edinburgh.

Doucherty, D. 1991. "sed & awk". O'Relly & Associates, Sebastopol, California.

Evans, R. and Gazdar, G. 1990. "The DATR Papers". Cognitive Science Research Paper CSRP-139, University of Sussex, Brighton.

Gale, W and Church, K.W. 1993. "Sentences in Bilingual Corpora". Computational Linguistics: 19(1), 75–102.

Goldfarb, C.F. 1990. The SGML Handbook Clarendon Press, Oxford.

Griswold, R.E. and M.T. Griswold, 1990. The Icon Programming Language (second edition). Prentice Hall, New Jersey.

Ide, N. and J. Véronis, (eds). 1995. The Text Encoding Initiative: Background and Context. Kluwer Academic Publishers, Dordrecht.

Manandhar, S. 1993. "CUF in Context". Computational Aspects of Constraint-Based Lingiuistics Description. ILIC/Department of Philosophy, University of Amsterdam.

Shieber, S., H. Uszkoreit, J. Robinson, and M. Tyson, 1983. "The formalism and implementation of PATR-II". Research on Interactive Acquisition and Use of Knowledge, 39–79. AI Center, SRI International, Menlo Park, Cal.

Sperberg-McQueen, C.M. and L. Burnard, (eds). 1994. Guidelines for Electronic text Encoding and Interchange. Chicago and Oxford.

Wall, L. and R.L. Schwartz, 1991. "Programming Perl". O'Reilly & Associates, Sebastopol, California.

Zajac, R. 1992. "Inheritance and Constraint-based Grammar Formalism". Computational Linguistics, 18(2).

The 'Terminology Market'

Christian Galinski

Infoterm
Heinestr. 18
A-1021 Wien
Tel.: +43 1 267 535 ext. 312
Fax: +43 1 216 3272

BEST COPY AVAILABLE

1. Terminology products and services for whom?

Terminologies emerge among others
- in science and technology in the course of scientific and technical development,
- in crafts and arts in the course of new techniques and skills,
- in society in conjunction with new conceptions and approaches.

They are created primarily by experts of various levels, in a multitude of subject-fields, and in an "evolutionary" rather than systematic way. The experts being the primary creators and users of their domain-specific terminologies also cause communication problems such as homonymy and polysemy, which some of them try to resolve by means of descriptive or prescriptive terminology work. Terminology work, therefore, is carried out in a large number of subject fields usually by groups of experts. In addition, it should be remembered, that it is a time-honoured scientific tradition to define what one is talking about in scientific and technical texts – a good tradition often neglected in scientific discourse today.

Since science and technology increasingly influence more and more ways of life and society, deficient terminologies are causing communication difficulties not only in the respective peer groups, but also create problems for many people who have to use specialised terminology
- at their work places,
- as consumers,
- as citizens, and

increasingly even in inter-family communication. Potentially and to a growing extend, everyone is or could become a user of any specialised terminology more or less frequently in his/her life.

The gradually emerging "terminology market" will offer terminological products and services – which in fact are a particular category of information products and services – to a variety of consumers and clients, such as
- terminology creators (e.g., researchers, administrators, etc.),
- terminology data producers (e.g., terminology database creators, specialised lexicographers, etc.),
- terminology data distributors (e.g., dictionary publishers, on-line information services, etc.),
- terminology users in general.

Terminology creators, data producers, and data distributors in most or many cases are also re-users of existing terminological data.

2. Terminology products

Terminology products comprise
- different kinds of terminological data in different forms for different purposes and different user groups,
- terminological tools for various purposes.

Terminological data (if terminology documentation is included) comprise three distinct types of data, viz.:
- terminological data proper (i.e., information on domain-specific concepts and their representation by linguistic and non-linguistic means supplemented by a variety of associated data),
- bibliographic data on a variety of different publications in the field of terminology,
- factual data on institutions, experts, programmes, and other activities in the field of terminology.

Each type of data requires a different type of information processing system. A comprehensive terminology information and documentation centre like Infoterm has to take care of all three types of data and further subdivide them into such categories (according to the respective "objects") for different purposes. The data as well as the respective software can be used as "products" and as basis for a variety of "services".

The volume of the above-mentioned types and categories of data may be estimated as follows:
- terminological data proper – about 50 million entries across all subject – fields (in some 200 languages which are of relevance or potential relevance in terminology) – the increase is about parallel to the increase of specialised knowledge,
- bibliographic data – about a 250 000 entries (of which an estimated 200 000 contain information on technical dictionaries and lexicons) – the annual increase can be estimated at about 10%,
- factual data – about 50 000 entries (80% of which concern terminology committees, commissions, and working groups as well as terminological institutions at international, regional, and national levels) – the increase is difficult to estimate, but the biggest problem is fluctuation!

2.1 Terminological data

Terminological data can be offered
- in conventionally published form,
- as an electronic publication (only the data as such in a given format or in combination with software)
- through on-line information services.

In palm-top computers or smaller pocket-size dictionaries the terminological data may be implemented in inseparable combination/integration with the respective software or even hardware.

Terminological data can be acquired by customers on the terminology market for internal use only or for re-use, in the course of terminology data interchange, etc., on a variety of different data carriers (floppy disk, CD-ROM, etc.). Different user groups need terminological data of different degrees of complexity for different purposes.

Therefore, it is highly economical to prepare multi-purpose terminological data for different purposes and users whose needs are taken care of by appropriately tailored customer-specific user-interfaces. Terminological data can also be used very efficiently as the intellectual "skeleton", or infrastructure, around which the contents of a domain-specific encyclopaedia can be organised.

2.2 Terminological tools

Terminology application software is the most common tool for the handling of terminological data in some way or other. **Terminology management** systems (TMS) are designed as dedicated tools to record, store, process, and output terminological data in a professional manner. There are different kinds of TMS for different purposes. **Terminology databases** consist of terminological data and a TMS to handle these data. **Terminology data banks** (TDB) are more or less sophisticated organisational structures established for the handling and maintenance of terminological data with the help of a TMS. TDBs can comprise several or many terminology databases.

TDBs are often supported by a TMS running on a mainframe, mini-computer, or work-station, whereas today most of the PC-based TMS are applied by individual users, small co-operatives (integrated or not by an appropriate LAN), or larger departments (where the individual work-places are usually linked by a more or less sophisticated LAN).

On the one hand, TMS are increasingly further developed into tools for various applications, such as
- computer-assisted translation,
- authoring,
- spare-part administration, etc.

On the other hand, TMS modules of varying degrees of sophistication are integrated into all kinds of application software.

In the future, appropriately designed TMS or TMS modules will find big markets particularly in applications, such as
- co-operative writing (today a high percentage of the citizens of developed countries work more or less intensively in some form or other as "technical writers"),
- documentation (in the meaning of information and documentation, and of archiving and filing), and
- co-operative terminology work.

If appropriate tools were available for computer-assisted terminology work, the preparation, processing, and maintenance of terminological data could be carried out faster, more efficiently, and according to modern quality approaches. Needless to mention that this would significantly aid the terminology market to develop.

3. Terminology services

At present the following terminology services already exist or are foreseeable in the future:
- consultancy and training services,
- outsourcing,
- information services.

3.1 Consultancy services and training

Consultancy services and training are most often needed with respect to application aspects, such as
- application of terminological principles and methods (especially the appropriate application of existing standards),
- selection and application of tools,
- terminology project management.

BEST COPY AVAILABLE

As a rule, experts today have not studied the basic theory of logic and epistemology underlying the science of sciences, or science theory which also comprises the basic theory of terminology. Therefore, they often need training in theoretical and methodological basics of terminology science and terminography. Large organisations/institutions often need to include terminological methods and tools into their information management or quality management schemes. Government agencies and other public authorities in many countries want to implement knowledge transfer policies, which would greatly benefit from the appropriate terminology planning methods. Institutions and organisations frequently also need advice with respect to legal problems (especially related to intellectual property rights) concerning the application of terminological data and tools.

However, it needs to be mentioned that with a few exceptions (e.g., China) these needs are still latent, decision makers not being aware of the usefulness and effectiveness of such services. For this reason and a lack of funds it is still a dormant market.

3.2 Outsourcing

Increasingly institutions and organisations of all sorts consider outsourcing an appropriate method to cope with certain limited terminological needs. For instance, outsourcing can refer to

- **research and development on demand, concerning new tools or applications, adaptation of existing tools, etc., such as**
- TDB design and implementation,
- meta-browsers for information networks, etc.
- **terminology work on demand with respect to**
  - terminology preparation,
  - terminology maintenance (including among others: revision and updating),
  - conversion or merging of terminological data,
  - evaluation of terminological data, etc.
- **maintenance and aftercare services with regard to**
  - TDB software maintenance and upgrading,
  - comprehensive data holdings maintenance, etc.

3.3 Information services

Increasingly, terminological products and services will – much like the general situation in the field of information and communication technology (ICT) – be available through all kinds of information services. They will also increasingly be integrated into other ICT applications.

4. Need for a dual terminology infrastructure in Europe

Given the amount of terminological entries across science, technology, and other subject-fields to be prepared in a multitude of languages, the monumental task of collecting terminology cannot be performed without the help of millions of experts who need to carry out this task anyway if they want to work and communicate efficiently. In most cases today such terminology work is carried out in the form of thousands of small efforts scattered all across the globe and subject-fields with little inter-connection. As a rule, it is performed in a non-commercial (not to say non-profit) framework.

Therefore, it would require a public or semi-public infrastructure

- to promote, organise, and co-ordinate terminological activities by domain experts in the public interest and with a non-profit and non-commercial approach which would take into account multiple user needs,
- to organise networking between many projects in which domain experts perform (preferably co-operatively) terminology work.

For the distribution of terminological data to different user groups with various user needs, efforts should be made to establish market-oriented networks for providing

- terminological data and other products as well as
- terminological services

on a commercial basis. The clients will thus have to pay for terminological products and services. The more clients can choose among an ever increasing variety of terminological products and services, the more affordable these will become.

Lexical Resources and Their Application

Martin Gellerstam

Göteborg University
Department of Swedish
S-412 98 Göteborg
Tel.: +46 31 77 34 544
Fax: +46 31 77 34 455
E-mail: gellerstam@svenska.gu.se

1. Lexical data as linguistic resources

Lexical data are valuable resources in a knowledge society. An American computer linguist, Martin Kay, in a paper about "The Dictionary of the Future and the Future of the Dictionary" (Kay 1984), compares big dictionaries with "Rolls Royce cars" and "country estates" . The metaphor may be a bit surprising, but it certainly takes a substantial amount of effort and skill to produce a good dictionary, and the ongoing discussions in computational linguistics about "reuse" of lexical data is a reflection of this fact.

Dictionary data have not always been considered as "resources" in the way this word has been commonly used up to now. The typical context for the word is "natural resources", things like water, timber, ore, etc. A widening of the concept to the field of linguistics came with computers and corpus linguistics. If you consult a printed dictionary to see how a word is spellt, you probably do not think of your dictionary as a resource; however if you use an automatic programme that spots all your misprints you start looking at lexical data in a new way. Thus, a spelling correction programme is just a start. We know that the dictionary has a greater potential than that. Let me just quote a leading lexicographer (Swanepoel 1994):

> The computer systems and tools that are becoming available both to the researcher, the practical lexicographer and the human user are opening up a myriad of possibilities for the presentation and utilization of masses of lexical information.

Therefore, it is not just a question of computer people handing over practical tools to lexicographers. Without lexical data, collected and systematized by skilled lexicographers, there would not be much to "reuse" by computational linguists. And the lexicon is a sine qua non for computational applications:

> The lexicon can be conceived as the point of conjunction of the different types of information to which any NLP system must have access: morphological, syntactic, semantic, pragmatic. (Calzolari 1989)

According to EU terminology, linguistic resources are divided into corpus resources, lexical resources, and tools. The borderline is not very distinct. It is a question of perpective if your resource is in horizontal order (textual data) or in vertical order (lexical data), and the same tool can be based on a tagged text or a lexicon. Furthermore, texual data form an important part of a dictionary (as sentence examples, etc). In fact, a corpus could be seen as an extension of the textual data actually used in a dictionary. In a CD-ROM version of a dictionary, the user could very well find a link to all the examples he or she could possibly need. This method is used already by dic-

tionary publishers. Thus, corpora are the stuff that dictionaries are made of, and dictionaries can refer to extended text reservoirs in the form of corpora.

So, what is the implication of the concept "lexical resources"? I would like to define it as "lexical data, preferably in machine-readable form, that can be used in lexical research and/or form the basis of commercial products". And "commercial products" could mean almost anything connected with words that is worth putting money into: dictionaries (computer-based or not), computer games, automatic hyphenation, spellchecking, writing aid, computer-aided learning, computer-aided translation, etc.)

2. Types of lexical data

As a starting point for our discussion of lexical resources, we will take a quick look at the lexical data involved. The following lexical facts – well-known to all lexicographers – reflect different aspects of lexical information necessary to describe the usage of a word. Owing to the tyranny of alphabetic order, dictionary publishers often take the opportunity to portion out the different types of lexical information into smaller dictionaries, specialized in one or two lexical aspects: dictionaries of spelling, pronunciation, definition, synonyms, idioms, etc. It remains to be seen if dictionaries stored in computer form will change this publishing tradition.

Words in a dictionary could be described according to the following principal aspects:

FORM	spelling
	pronunciation
	inflexion
	word class
MEANING	definition/equivalent
	synonyms (antonyms, hyperonyms, etc)
	thesaurus classification
CONTEXT	grammatical collocations
	lexical collocations
	idioms
	valency
PRAGMATICS	distribution (domain, register, style)
	frequency

If you look at these lexical categories – most of which have a published dictionary counterpart – you will find that they are more or less suited for

computational application. The bestsellers today are computer programmes that check your spelling or your grammar and style and give you information about synonyms. And for these programmes to be able to handle your text, they have to have access to inflexion and word class. Other categories (like valency and collocations) will be of great importance in automatic text analysis. Frequency could also be used in this context.

3. Carriers of lexical data

Lexical data will reach the language user in a variety of machine-readable dictionaries and computer programmes, ranging from simple spelling-checking devices to sophisticated products of computational technology. The following categories reflect a hierarchy from simple carriers of lexical data to more complex ones. The hierachy also reflects the degree of explicitness: from dictionaries where the human reader can pick the details for his or her understanding of a text to formalized lexicons where the information must be explicit.

1. Word frequency lists (word form -> lemma)
2. Printed dictionaries in machine-readable form
3. Printed dictionaries in machine-readable form with linguistic codes and classification; also recently published CD-ROM versions
4. Machine lexicons, classified, encoded, and with selected information (often designed for automatic lemmatization)
5. Lexical data bases
6. Computational linguistics lexicons

Word frequency lists were produced as a result of many frequency counts in the sixties and seventies. One example is the American Brown Corpus (Kučera & Francis 1967) which has had a marked influence on later corpus investigations, especially as a model for corpus collection.

Printed dictionaries in machine-readable form can also be dated back to the sixties when the Merriam-Webster Seventh New Collegiate Dictionary was put into machine-readable form by J. Olney and his colleagues.

Printed dictionaries with explicit linguistic codes and classification have been published since the first edition of the Longman Dictionary (LLDOCE 1981). Later editions of both this dictionary and Collins COBUILD dictionary have lately appeared also in CD-ROM form.

Thesaurus dictionaries date back to Roget's Thesaurus from the 19th century. Later efforts in the same direction have been published as pedagogical

variants of regular dictionaries, from Longman's Lexicon of Contemporary English (1981) to Longman's Language Activator (1993).

Machine lexicons are not designed to be read by humans but provide explicit lexical information for performing specific tasks, e.g., automatic lemmatization. The words of a text are confronted with a list of forms listed in the dictionary. If a certain form is found in the text, the word is associated to the correct lemma (including part-of-speech).

Lexical data bases (LDBs) contain formalized information at many descriptive levels. It is one of the chief tools today for processing great quantities of lexical information. It can be used for various types of linguistic applications and for general research in the lexical field. A data base management system provides the user with tools which enable him to access the data without necessarily being familiar with the internal or physical organisation, but only with the type of information he can retrieve.

Computational linguistic lexicons are more complex tools for parsing, for artificial intelligence (question-answering) and for Machine Translation.

3. TELRI resources

A quick look at the TELRI Language Resources gives an impressive overview of lexical data from a wide range of European languages. Efforts to bring all of this data together and make it accessible – or at least make a catalogue accessible – should be a priority for TELRI. We certainly need a European counterpart to the Linguistic Data Consortium in the United States.

It is difficult to avoid a certain disagreement about lexical terminology. When you find a "database with morphological information" among the TELRI lexical data you do not know if you are confronted with a well-structured database containing various kinds of information about inflection, derivation, etc. or perhaps a machine lexicon for lemmatization. To be able to compare the different types of information among the TELRI countries, you should not really compare a database with inflection categories with a lemmatizer. To have a database does not imply that you have all relevant lexical information: it just means that you have stored the lexical facts in a certain form with multiple access. If you have a "phraseological dictionary", how does this relate to things like "collocation tools" (which is not a lexical resource), etc.?

After these reservations, the lexical resources situation could be summed up in the following way:

- There are quite a number of MRDs but few fullfledged databases with a variety of lexical information
- The majority of TELRI members have tackled the crucial question of lemmatization which often – but not always – correlates with the existence of spelling-checkers
- Semantic information is scarce: there are just a few dictionaries of synonyms and semantic tagging (e.g., tagging of definitions), and thesauruses are rare.

4. Areas of Application

To ask for the application of lexical resources is like asking for linguistic application in general. If the lexicon – at least in principle – is "the point of conjunction of the different types of information to which any NLP system must have access" (see above), you will find it difficult to say what is not an application of lexical resources. In this context, however, I will just point out a few obvious fields of application.

The first area that should be mentioned is the field of lexical research itself, which is not only producing applications but which is also an activity putting existing applications to the test. Such applications range from simple concordances and tagged texts to lexical databases and alignment methods (notice the fuzzy borderline between "lexical data", "textual data", and "tools").

Lexical resources are also used in various other fields of research, e.g., *psychology*, where lexical data is needed in fields like language learning, testing of patients with brain diseases, etc. In sociology, vocabulary data are used to reflect cultural and ideologic development in society.

Lexicographic practice is a field where lexical applications are put to continuous test. Applications cover the whole dictionary production line from corpus collection over the harbouring of lexical facts in a database with multiple access to the final production of a dictionary article. However, the one outstanding lexical application is the final dictionary itself – on paper, diskettes, CD-ROM, etc.

Even if the dictionary itself is the image of lexical resources, writing aids of different kinds are the most typical computational commercial product. To begin from basics, there are general text checkers that check practical things like starting a new sentence with a capital letter, spotting extra spaces between words, etc. *Spelling checkers* are usually based on a collection of wordforms representing an actual corpus or a list of wordforms generated from a dictionary. Text verification to find spelling errors and for automatic

hyphenation is probably the number one commercial application. Spelling checking is a relatively easy task for a language like English with little morphology but becames a more complex task in a language with rich morphology. Spelling checking facilities are more or less standard ingredients in word processing today. This is also true of synonymy information (sometimes advertized to give you an impression that languages contain hundreds of thousands of synonyms). *Style checkers* have developed from the first simple spelling correction systems. Modern style checkers include checking of particular words from stylistic point of view ("why do you use the passive form?"), parsers for spotting grammatical errors (like congruence), and checking of contextual data ("have you used the right preposition after the verb?").

A more specialized type of writing aid is *computer-aided translation (CAT)*, where a computer programme looks up words and phrases and suggests translations to the human translator.

Language learning applications based on lexical data can be used in various types of interactive teaching of written language skills. Computer programmes can assist in tasks like sentence restructuring, checking of translation and dictation tasks, cloze testing (filling in omitted words in a text), and dictionary look-up.

Other fields of knowledge where lexical data can be applied is *information retrieval* (this is where the thesaurus comes in), various kinds of applications in *artificial intelligence* such as question-answering. The need for a comprehensive lexicon for *machine translation (MT)* is widely acknowledged today.

5. Final remarks

In this brave new world of possibilities and application of lexical resources, it may be necessary to add a few words of warning concerning the general handling of lexical data. Today, great efforts are put into *standardization of texts* and lexical data in the framework of the Text Encoding Initiative. The idea is of course that a home-made formalism for your lexical data is less than practical in a society where this sector is growing steadily and exchange of data is becoming more and more frequent. On the other hand, standardization is a laborious and expensive task – especially if we are talking about standardization of already existing corpora – and you may not quickly reap the full benefit of your effort. You must make a rational weighing-up of pros and cons.

Another problem is the *dissemination* of your lexical application. Be sure about the legal and economic implications before you offer your data free of

charge or for a sum of money, for scholarly or commercial use. How many public domain dictionaries have you come across when surfing on Internet? Also, do not forget to protect your *copyright* in dealing with commercial partners.

●

References

Calzolari, Nicoletta. 1989. "Computer-Aided Lexicography: Dictionaries and Word Data Bases". In: Computational Linguistics. An International Handbook on Computer Oriented Language Research and Applications. Berlin & New York: Walter de Gruyter.

Collins COBUILD English Language Dictionary. 1987. London and Glasgow: Collins.

Kay, Martin. 1983. "The Dictionary of the Future and the Future of the Dictionary". In: Linguistica computazionale, 3 (1983).

Kučera, Henry & Francis, W. Nelson. 1967. Computational Analysis of Present-Day American English. Providence, Rhode Island: Brown University Press

Longman Dictionary of Contemporary English (LLDOCE). 1981. Harlowe and London: Longman Group Limited.

Longman's Lexicon of Contemporary English. 1981. Harlowe: Longman Group Limited

Longman Language Activator. 1993. Harlowe: Longman Group UK Limited

Swanepoel, Piet. 1994. "Problems, Theories and Methodologies in Current Lexicographic Semantic Research". In: Willy Martin & Willem Meijs & Margreet Moerland & Elsemiek ten Pas & Piet van Sterkenburg & Piek Vossen (eds), Euralex 1994 Proceedings. Amsterdam 1994, p. 11–26.

Encoding Standards for Linguistic Corpora

Nancy Ide

Department of Computer Science
Vassar College
Poughkeepsie, New York 12601 (U.S.A.)
and
Laboratoire Parole et Langage
CNRS/Université de Provence
29, Avenue Robert Schuman
13621 Aix-en-Provence Cedex 1 (France)
Tel: +1 914 437 5988
Fax: +1 914 437 7187
E-mail: ide@cs.vassar.edu/ide@univ-aix.fr

1. Introduction

The past few years have seen a burst of activity in the development of statistical methods which, applied to massive text data, have in turn enabled the development of increasingly comprehensive and robust models of language structure and use. Such models are increasingly recognized as an invaluable resource for natural language processing (NLP) tasks, including machine translation.

The upsurge of interest in empirical methods for language modelling has led inevitably to a need for massive collections of texts of all kinds, including text collections which span genre, register, spoken and written data etc., as well as domain- or application-specific collections, and, especially, multilingual collections with parallel translations. In the latter half of the 1980's, very few appropriate or adequately large text collections existed for use in computational linguistics research, especially for languages other than English. Consequently, several efforts to collect and disseminate large mono- and multi-lingual text collections have been recently established, including the ACL Data Collection Initiative (ACL/DCI), the European Corpus Initiative (ECI), the U. S. Linguistic Data Consortium (LDC), MULTEXT in Europe, etc. It is widely recognized that such efforts constitute only a beginning for the necessary data collection and dissemination efforts, and that considerable work to develop adequately large and appropriately constituted textual resources still remains.

The demand for extensive reusability of large text collections in turn requires the development of standardized encoding formats for this data. It is no longer realistic to distribute data in ad hoc formats, since the effort and resources required to clean up and reformat the data for local use is at best costly, and in many cases prohibitive. Because much existing and potentially available data was originally formatted for the purposes of printing, the information explicitly represented in the encoding concerns a particular physical realization of a text rather than its logical structure (which is of greater interest for most NLP applications), and the correspondence between the two is often difficult or impossible to establish without substantial work. Further, as data become more and more available and the use of large text collections become more central to NLP research, general and publicly available software to manipulate the texts is being developed which, to be itself reusable, also requires the existence of a standard encoding format.

A standard encoding format adequate for representing textual data for NLP research must be (1) capable of representing the different kinds of information across the spectrum of text types and languages potentially of interest to the

NLP research community, including prose, technical documents, newspapers, verse, drama, letters, dictionaries, lexicons, etc.; (2) capable of representing different levels of information, including not only physical characteristics and logical structure (as well as other more complex phenomena such as intra- and inter-textual references, alignment of parallel elements, etc.), but also interpretive or analytic annotation which may be added to the data (for example, markup for part of speech, syntactic structure, etc.); (3) application independent, that is, it must provide the required flexibility and generality to enable, possibly simultaneously, the explicit encoding of potentially disparate types of information within the same text, as well as accommodate all potential types of processing. The development of such a suitably flexible and comprehensive encoding system is a substantial intellectual task, demanding (just to start) the development of suitably complex models for the various text types as well as an overall model of text and an architecture for the encoding scheme that is to embody it.

2. The Text Encoding Initiative

In 1988, the Text Encoding Initiative (TEI) was established as an international co-operative research project to develop a general and flexible set of guidelines for the preparation and interchange of electronic texts. The TEI is jointly sponsored by the Association for Computers and the Humanities, the Association for Computational Linguistics, and the Association for Literary and Linguistic Computing. The project has had major support from the U. S. National Endowment for the Humanities (NEH), Directorate XIII of the Commission of the European Communities (CEC/DG-XIII), the Andrew W. Mellon Foundation, and the Social Science and Humanities Research Council of Canada.

In May 1994, the TEI issued its Guidelines for the Encoding and Interchange of Machine-Readable Texts, which provide standardized encoding conventions for a large range of text types and features relevant for a broad range of applications, including natural language processing, information retrieval, hypertext, electronic publishing, various forms of literary and historical analysis, lexicography, etc. The Guidelines are intended to apply to texts, written or spoken, in any natural language, of any date, in any genre or text type, without restriction on form or content. They treat both continuous materials (running text) and discontinuous materials such as dictionaries and linguistic corpora. As such, the TEI Guidelines answer the fundamental needs of a wide range of users: researchers in computational linguistics, the

humanities, sciences, and social sciences; publishers; librarians and those concerned generally with document retrieval and storage; as well as the growing language technology community, which is amassing substantial multi-lingual, multi-modal corpora of spoken and written texts and lexicons in order to advance research in human language understanding, production, and translation.

The rules and recommendations made in the TEI Guidelines conform to the ISO8879, which defines the Standard Generalized Markup Language, and ISO 646, which defines a standard seven-bit character set in terms of which their commendations on character-level interchange are formulated.* SGML is an increasingly widely recognized international markup standard which has been adopted by the US Department of Defense, the Commission of European Communities, and numerous publishers and holders of large public databases.

2.1 Overview

Prior to the establishment of the TEI, most projects involving the capture and electronic representation of texts and other linguistic data developed their own encoding schemes, which usually could only be used for the data for which they were designed. In many cases, there had been no prior analysis of the required categories and features and the relations among them for a given text type, in the light of real and potential processing and analytic needs. The TEI has motivated and accomplished the substantial intellectual task of completing this analysis for a large number of text types, and provides encoding conventions based upon it for describing the physical and logical structure of many classes of texts, as well as features particular to a given text type or not conventionally represented in typography. The TEI Guidelines also cover common text encoding problems, including intra- and inter-textual cross reference, demarcation of arbitrary text segments, alignment of parallel elements, overlapping hierarchies, etc. In addition, they provide conventions for linking texts to acoustic and visual data.

The TEI's specific achievements include:

* For more extensive discussion of the project's history, rationale, and design principles see TEI internal documents EDP1 and EDP2 (available from the TEI) and (Ide and Sperberg-McQueen 1995: 5—15) and (Sperberg-McQueen and Burnard 1995: 17—39), both published in a special triple issue on the TEI in *Computers and the Humanities*.

1. a determination that the Standard Generalized Markup Language (SGML) is the framework for development of the Guidelines;
2. the specification of restrictions on and recommendations for SGML use that best serves the needs of interchange, as well as enables maximal generality and flexibility in order to serve the widest possible range of research, development, and application needs;
3. analysis and identification of categories and features for encoding textual data, at many levels of detail;
4. specification of a set of general text structure definitions that is effective, flexible, and extensible;
5. specification of a method for in-file documentation of electronic texts compatible with library cataloguing conventions, which can be used to trace the history of the texts and thus assist in authenticating their provenance and the modifications they have undergone;
6. specification of encoding conventions for special kinds of texts or text features, including:

    a. character sets
    b. language corpora
    c. general linguistics
    d. dictionaries
    e. terminological data
    f. spoken texts
    g. hypermedia
    h. literary prose
    i. verse
    j. drama
    k. historical source materials
    l. text critical apparatus

3. Basic architecture of the TEI scheme

3.1 General architecture

The TEI Guidelines are built on the assumption that there is a common core of textual features shared by virtually all texts, beyond which many different elements can be encoded. Therefore, the Guidelines provide an extensible framework containing a common core of features, a choice of frameworks or bases, and a wide variety of optional additions for specific

applications or text types. The encoding process is seen as incremental, so that additional markup may be easily inserted in the text.

Because the TEI is an SGML application, a TEI conferment document must be described by a document type definition (DTD), which defines tags and provides a BNF grammar description of the allowed structural relationships among them. A TEI DTD is composed of the core tagsets, a single base tag set, and any number of user selected additional tagsets, built up according to a set of rules documented in the TEI Guidelines.

At the highest level, all TEI documents conform to a common model. The basic unit is a text, that is, any single documenter stretch of natural language regarded as a self-contained unit for processing purposes. The association of such a unit with a header describing it as a bibliographic entity is regarded as a single TEI element. Two variations on this basic structure are defined: a collection of TEI elements, or a variety of composite texts. The first is appropriate for large disparate collections of independent texts, for example in language corpora, or collections of unrelated papers in an archive; the second applies to cases such as the complete works of a given author, which might be regarded simultaneously as a single text in its own right and as a series of independent texts.

Often, it is necessary to encode more than one view of a text — for example, the physical and the linguistic or the formal and the rhetorical. One of the essential features of the TEI Guidelines is that they offer the possibility to encode many different views of a text, simultaneously if necessary. A disadvantage of SGML is that it uses a document model consisting of a single hierarchical structure; often, different views of a text define multiple, possibly overlapping hierarchies (for example, the physical view of a print version of a text, consisting of pages sub-divided into physical lines, and the logical view consisting of, say, paragraphs sub-divided into sentences)which are not readily accommodated by SGML's document model. The TEI has identified several possible solutions to this problem in addition to SGML's concurrent structures mechanism, which, because of the processing complexity it involves, is not a thoroughly satisfactory alternative.

The TEI Guidelines provide sophisticated mechanisms for linking and alignment of elements, both within a given text and between texts, as well as links to data not in the form of ASCII text such as sound and images. Much of the TEI work on linkage was accomplished in collaboration with those working on the Hypermedia/Time-based Document Structuring Language (HyTime), recently adopted as an SGML-based international standard for hypermedia structures.

3.2 The TEI base tagsets

Eight distinct TEI base tagsets are proposed:

1. prose
2. verse
3. drama
4. transcribed speech
5. letters or memos
6. dictionary entries
7. terminological entries
8. language corpora and collections

The first seven are intended for documents which are predominantly composed of one type of text; the last is provided for use with texts which combine these basic tagsets. Additional base tagsets will be provided in the future.

Each TEI base tagset determines the basic structure of all the documents with which it is to be used. More exactly, it defines the components of text elements, combined as described above. In practice, so far, almost all the TEI bases defined are similar in their basic structure, though they can vary if necessary. However, they differ in their components: for example, the kind of sub-elements likely to appear within the divisions of a dictionary will be entirely different from those likely to appear within the divisions of a letter or a novel. To accommodate this variety, the constituents of all divisions of a TEI text element are not defined explicitly, but in terms of SGML parameter entities, which behave similar to a variable declaration in a programming language; the effect of using them here is that each base tag set can provide its own specific definition for the constituents of texts, which can, moreover, be modified by the user.

3.3 The core tagsets

Two core tagsets are available to all TEI documents unless explicitly disabled. The first defines a large number of elements which may appear in any kind of document-coinciding more or less with that set of discipline-independent textual features concerning which consensus has been reached. The second defines the header, providing something analogous to an electronic title page for the electronic text.

The core tagset common to all TEI bases provides means of encoding with a reasonable degree of sophistication the following list of textual features:

1. Paragraphs.
2. Segmentation, for example into orthographic sentences.
3. Lists of various kinds, including glossaries and indexes
4. Typographically highlighted phrases, whether unqualified or used to mark linguistic emphasis, foreign words, titles etc.
5. Quoted phrases, distinguishing direct speech, quotation, terms and glosses, cited phrases etc.
6. Names, numbers and measures, dates and times, and similar data-like phrases.
7. Basic editorial changes (e.g., correction of apparent errors; regularization and normalization; additions, deletions and omissions)
8. Simple links and cross references, providing basic hypertextual features.
9. Pre-existing or generated annotation and indexing
10. Passages of verse or drama, distinguishing for example speakers, stage directions, verse lines, stanzaic units, etc.
11. Bibliographic citations, adequate for most commonly used bibliographic packages, in either a free or a tightly structured format
12. Simple or complex referencing systems, not necessarily dependent on the existing SGML structure.

There are few documents which do not exhibit some of these features, and none of these features is particularly restricted to any one kind of document. In most cases, additional more specialized tagsets are provided to encode aspects of these features in more detail, but the elements defined in this core should be adequate for most applications most of the time.

Features are categorized within the TEI scheme based on shared attributes. The TEI encoding scheme also uses a classification system based upon structural properties of the elements, that is, their position within the SGML document structure. Elements which can appear at the same position within a document are regarded as forming a model class: for example, the class phrase includes all elements which can appear within paragraphs but not spanning them, the class chunk includes all elements which cannot appear within paragraphs (e. g. , paragraphs), etc. A class inter is also defined for elements such as lists, which can appear either within or between chunk elements.

Classes may have super- and sub-classes, and properties (notably, associated attributes) may be inherited. For example, reflecting the needs of many TEI users to treat texts both as documents and as input to databases, a sub-class of phrase called data is defined to include data-like features such as names of

persons, places or organizations, numbers and dates, abbreviations and measures. The formal definition of classes in the SGML syntax used to express the TEI scheme makes it possible for users of the scheme to extend it in a simple and controlled way: new elements may be added into existing classes, and existing elements renamed or undefined, without any need for extensive revision of the TEI document type definitions.

3.4 The TEI header

The TEI header is believed to be the first systematic attempt to provide in-file documentation of electronic texts. The TEI header allows for the definition of a full AACR2-compatible bibliographic description for the electronic text, covering all of the following:

1. the electronic document itself
2. sources from which the document was derived
3. encoding system
4. revision history

The TEI header allows for a large amount of structured or unstructured information under the above headings, including both traditional bibliographic material which can be directly translated into an equivalent MARC catalogue record, as well as descriptive information such as the languages it uses and the situation within which it was produced, expansions or formal definitions for any code books used in analyzing the text, the setting and identity of participants within it, etc. The amount of encoding in a header depends both on the nature and the intended use of the text. At one extreme, an encoder may provide only a bibliographic identification of the text. At the other, encoder swishing to ensure that their texts can be used for the widest range of applications can provide a level of detailed documentation approximating to the kind most often supplied in the form of a manual.

A collection of TEI headers can also be regarded as a distinct document, and an auxiliary DTD is provided to support interchange of headers alone, for example, between libraries or archives.

3.5 Additional tagsets

A number of optional additional tagsets are defined by the Guidelines, including tagsets for special application areas such as alignment and linkage of text segments to form hypertexts; a wide range of other analytic elements and attributes; a tagset for detailed manuscript transcription and another for the

recording of an electronic variorum modelled on the traditional critical apparatus; tagsets for the detailed encoding of names and dates; abstractions such as networks, graphs or trees; mathematical formulae and tables, etc.

In addition to these application-specific specialized tagsets, a general purpose tagset based on feature structure notation is proposed for the encoding of entirely abstract interpretations of a text, either in parallel or embedded within it. Using this mechanism, encoders can define arbitrarily complex bundles or sets of features identified in a text. The syntax defined by the Guidelines formalizes the way in which such features are encoded and provides for a detailed specification of legal feature value/pair combinations and rules (a feature system declaration) determining, for example, the implication of under-specified or defaulted features. A related set of additional elements is also provided for the encoding of degrees of uncertainty or ambiguity in the encoding of a text.

A user of the TEI scheme may combine as many or as few additional tagsets as suit his or her needs. The existence of tagsets for particular application areas in the Guidelines reflects, to some extent, accidents of history: no claim to systematic or encyclopedic coverage is implied. It is expected that new tagsets will be defined as a part of the continued work of the TEI and in related projects.*

4. Information about the TEI

The TEI Guidelines for Electronic Text Encoding and Interchange are available as follows:

in paper (1300 pp., 2 volumes), at a cost of $75 US or 50 pounds sterling, sent to:

TEI Orders
Oxford University Computing Services
13 Banbury Road
Oxford OX2 6NN

* For example, the European project MULTEXT, in collaboration wit EAGLES, is developing a specialized Corpus Encoding Standard for NPL applications based on the TEI Guidelines.

electronically via the World Wide Web at the following sites:

http://www-tei.uic.edu/orgs/tei
http://etext.virginia.edu/TEI.html

electronically via anonymous ftp from any of the following:

ftp-tei.uic.edu (in pub/tei and its subdirectories)
sgml1.ex.ac.uk (in tei/p3 and its subdirectories)
ftp.ifi.uio.no (in pub/SGML/TEI)

electronically informatted ASCII-only via Listserv, by sending electronic mail to listserv @uicvm.uic.edu containing the following line:

get p3ascii package

The TEI also maintains a publicly-accessible List Serv list, TEI-L, housed at the University of Illinois at Chicago. To subscribe, send electronic mail to **listserv @uicvm.uic.edu** containing the text **Subscribe TEI-L J. Q. Public** (substitute your name for "J. Q. Public")

Additional information can be obtained by contacting one of the TEI editors:

C. M. Sperberg-McQueen
University of Illinois at Chicago (M/C 135)
Computer Center
1940 W. Taylor St.
Chicago, Illinois 60612-7352 US
E-mail: u35395@uicvm.uic.edu
tel: +1 (312) 413-0317
fax: +1 (312) 996-6834

Lou Burnard
Oxford University Computing Service
13 Banbury Road
Oxford OX26NN
United Kingdom
E-mail: lou@vax.ox.ac.uk
tel: +44 (865) 273200
fax: +44 (865) 273275

Acknowledgments – The TEI has been funded by the U.S. National Endowment for the Humanities (NEH), Directorate XIII of the Commission of the European Communities (CEC/DG-XIII), the Andrew W. Mellon Foundation, and the Social Science and Humanities Research Council of Canada. Some material in this paper has been adapted from other TEI documents written by the TEI editors Michael Sperberg-McQueen and Lou Burnard, and chairs and members of various TEI committees.

References

Bryan, M. 1988. SGML: An Author's Guide, New York: Addison-Wesley.

Coombs, J. H., Renear, A. H., and DeRose, S. J. 1987. "Markup systems and the future of scholarly text processing". Communications of the ACM 30(11): 933–947.

Goldfarb, C. F. 1990. The SGML Handbook. Oxford: Clarendon Press.

Ide, N., Sperberg-McQueen, C. M. 1995. "The Text Encoding Initiative: Its History, Goals, and Future Development". Computers and the Humanities (Special Issue on the Text Encoding Initiative) 29(1): 5–15.

Ide, N., Véronis, J. (eds.) 1995. Computers and the Humanities (Special Issue on the Text Encoding Initiative) 29(1–3).

International Organization for Standards 1986. ISO 8879: Information Processing-Text and Office Systems-Standard Generalized Markup Language (SGML). Geneva: ISO.

International Organization for Standards 1992. ISO/IEC DIS 10744: Hypermedia/Time-based Document Structuring Language (Hytime). Geneva:ISO.

Sperberg-McQueen, C. M., Burnard, L. 1995. "The Design of the TEI Encoding Scheme". Computers and the Humanities (Special Issue on the Text Encoding Initiative) 29(1): 17–39.

Sperberg-McQueen, C. M., Burnard, L. 1994. Guidelines for Electronic Text Encoding and Interchange. Chicago and Oxford: Text Encoding Initiative.

van Herwijnen, E. 1991. Practical SGML. Dordrecht: Kluwer Academic Publishers.

Appendix: TEI Guidelines Table of Contents

Part I: Introduction
1. About These Guidelines
2. Concise Summary of SGML
3. Structure of the TEI Document Type Declarations

Part II: Core Tags and General Rules
4. Characters and Character Sets
5. The TEI Header
6. Tags Available in All TEI DTDs

Part III: Base Tagsets
7. Base Tagset for Prose
8. Base Tagset for Verse
9. Base Tagset for Drama
10. Base Tagset for Transcriptions of Spoken Texts
11. Base Tagset for Letters and Memoranda
12. Base Tagset for Printed Dictionaries
13. Base Tagset for Terminological Data
14. Base Tagset for Language Corpora and Collections
15. User-Defined Base Tagsets

Part IV: Additional Tagsets
16. Segmentation and Alignment
17. Simple Analytic Mechanisms
18. Feature Structure Analysis
19. Certainty
20. Manuscripts, Analytic Bibliography, and Physical Description of the source Text
21. Text Criticism and Apparatus
22. Additional Tags for Names and Dates
23. Graphs, Digraphs, and Trees
24. Graphics, Figures, and Illustrations
25. Formulae and Tables
26. Additional Tags for the TEI Header

Part V: Auxiliary Document Types
27. Structured Header
28. Writing System Declaration

29. Feature System Declaration
30. Tagset Declaration

Part VI: Technical Topics
31. TEI Conformance
32. Modifying TEI DTDs
33. Local Installation and Support of TEI Markup
34. Use of TEI Encoding Scheme in Interchange
35. Relationship of TEI to Other Standards
36. Markup for Non-Hierarchical Phenomena
37. Algorithm for Recognizing Canonical References

Part VII: Alphabetical Reference List of Tags and Classes

Part VIII: Reference Material
38. Full TEI Document Type Declarations
39. Standard Writing System Declarations
40. Feature System Declaration for Basic Grammatical Annotation
41. Sample Tagset Declaration
42. Formal Grammar for the TEI-Interchange-Format Subset of SGML

Machine Translation: State of the Art, Trends and the User Perspective

Steven Krauwer

Research Institute for Language and Speech (OTS)
Utrecht University
Trans 10
3512 JK Utrecht
The Netherlands
Tel. +31 30 253 6050
Fax: +31 30 253 6000
E-mail: steven.krauwer@let.ruu.nl

1. Machine Translation, does it exist?

The only honest answer to this question is NO. If we define a Machine Translation system as a system capable of successfully imitating the behaviour of a human translator, we still have a long way to go. As a matter of fact, we do not even know what successfully imitating means: there exists no formal or even semi-formal description of what an MT system is supposed to do. But even in the absence of such a formal description, we know from direct observation that no computer program is capable of delivering translations even remotely approaching the quality of translations made by professional human translators.

The biggest problem in practice is the disambiguation problem. Human beings have no problems in picking out word senses and assigning the proper interpretations and translations to phrases, sentences, and texts, whereas even the most sophisticated computer system would have a hard time identifying the proper referent of the word "they" in the following two sentences (the examples are borrowed from Jerry Hobbs):

(1) The policemen fired at the demonstrating students because they wanted revolution.

(2) The policemen fired at the demonstrating students because they feared revolution.

The reason why we can easily interpret both sentences correctly is that we have more than just our linguistic knowledge to rely on, and it is only when one tries to analyze texts on the basis of linguistic rules only that it becomes clear that more often than not extralinguistic knowledge is needed to assign the appropriate analysis. The real problem is of course not that we could not store somewhere the information that students are more likely to want revolution than policemen and that policemen are supposed to stop revolutions from happening, but rather that it is totally unpredictable what knowledge about our culture, our behaviour, our technology, etc. will be needed in order to disambiguate texts to be translated.

But although it is very tempting to stick to the "NO" answer to the question about the existence of MT and just give up the enterprise (at least for the time being), it somehow feels unsatisfactory to go for this option. After all, the idea that a MT system should simulate the behaviour of a human translator is based on a rather arbitrary interpretation of what MT could be.

If we focus on the underlying problem (language barriers) rather than the instrument traditionally used to solve it (translators) and define an MT system as a system capable of reducing the negative effects of these language barriers, we can identify various types of existing systems which do help us to reduce

these problems, even if they are very poor imitations of the human translator or do not even pretend to be an imitation (and the well-known comparison between birds and aeroplanes comes to mind immediately).

2. Users of MT systems

If we assume that the purpose of MT is to reduce the problems caused by the fact that different people speak different languages, we have to accept that "reduction" means different things to different people. For some people, reduction may refer to the cost of translation; for others, speed may play a more important role, and yet others may want to gain better access to information made available in foreign languages.

Therefore, we have to distinguish a number of classes of users of MT systems. For the purpose of this paper, a distinction into three classes will be made, but it should be kept in mind that more fine-grained classifications are possible.

2.1 Big companies and institutions

Many big companies and institutions operating on an international scale are faced with a huge translation load which justifies an in-house translation service. In such an environment, MT may become an attractive option if at least one of the following criteria is met:

- the translation quality is increased,
- turn-around times are improved, or
- the translation cost per unit is reduced.

If we look at what MT currently has to offer, we get the following picture on the negative side:

- the quality of full MT system output is generally poor, i.e., (post-editing is necessary);
- introduction of MT requires a high initial investment in training of people, acquisition of hardware and software, and in the customisation of dictionaries;
- it requires a different organisation of the work flow;
- the cost of computational support for such systems is high in comparison with e.g., translators using word processors on PCs.

On the positive side we have:

- high speed;
- relatively low exploitation cost in comparison with human translators.

If the volume is big enough, the speed increase or the cost reduction may be significant. Just to mention an example: the system Logos was used by Lexi-Tech to translate 500 000 pages of user documentation for 12 Canadian patrol frigates from English into French. If we assume (optimistically) that a human translator can manage some 8 pages of text per day, this job will require an effort of more than 300 man years. If the price per translated page is estimated at 50 ecu, the overall translation cost will be in the order of 25 Mecu if done by human translators. It is clear that under such circumstances the use of MT may mean the difference between practically and financially feasible and unfeasible. Even if an MT system produces raw translations which require significant human post-editing, an prospective overall cost reduction of, say, 20% will justify a considerable initial investment.

Our conclusion is that this approach will work well for large volumes. If the volume is not large enough, there is little chance that there will be any significant return on the investment.

Typical examples of translation systems used by big companies or organisations:

- Systran, used by the European Commission;
- Metal, used by the Union Bank of Switzerland to translate information technology and telecommunications texts from German into English, for internal use;
- Kielikone, used by Nokia to translate telecommunications customers documentation from Finnish into English.

2.2 Professional and occasional translators

Professional translators are in a different position, since translation is their core business and their specific area of expertise. They often work on a freelance basis or in small companies. Their annual turnover does not usually justify major investments, and they do not normally want their jobs to be taken over completely by their computers. Occasional translators have even less reason to invest in expensive MT systems.

The main requirements of professional and occasional translators are:

- increase of productivity,
- better quality, and
- low cost (investments and operation).

What MT has to offer on the negative side:

- poor quality on big, expensive systems;
- very poor quality on cheap PC based systems;

but on the positive side we have:
- editing and checking tools;
- dictionaries and term banks;
- translation memories, i.e., systems capable of storing fragments of texts and their translations, with retrieval capabilities based on exact or fuzzy pattern matching.

The conclusion is that there is little reason for this group to resort to MT systems. The big ones are too expensive, and the cheap ones deliver a translation quality which requires so much editing that most translators would prefer to do the translations themselves rather than correcting the system's results.

Yet, in the area of tools and facilities, there are a number of products (already on the market) which help the translator to increase their productivity (e.g., translation memories, which help to avoid translating the same or similar fragments more than once) and the quality of their work (e.g., by helping to make consistent use of the customer's standard terminology).

Some typical examples
- Mono- and bilingual dictionaries on CD-ROM;
- Translator's Workbench 2 (Trados, Germany), including translation memory, terminology database, translation editor, terminology extraction, etc;
- Translator Work Station (CITI, Canada), including access to previously translated texts, word processing, terminology management, document comparison, etc.

2.3. Monolingual users

The third group we can identify are those users who want to be able to produce translations from or into languages they do not know themselves.

Their requirements are the following:
- low cost;
- easy to operate;
- translations should be usable as they are (i.e., no checking or post-editing).

What MT has to offer on the negative side:
- poor quality on cheap systems, not good enough to send out to third parties.

On the positive side:
- possibly good enough for internal use (e.g., relevance checking of incoming messages);
- modem or network access to professional translation services (which may or may not use MT).

Conclusion: MT is no real option for monolingual users if they need translations for communication with other parties (e.g., user documentation of their products, business letters), but network access to translation services may offer a solution. Most modern PC based translation systems offer good opportunities for relevance checking of incoming information, provided it is available in electronic form.

Typical examples are:
- PC based systems such as GlobaLink, MicroTac;
- Access to Systran via Minitel in France.

3. Current trends on the market

A number of interesting developments can be seen on the market, all of which will contribute to overcoming the language barriers in Europe:

(1) Big MT systems are becoming more cost-effective because of
   - better environments;
   - better integration of MT in, e.g., document production;
   - cheaper but more powerful hardware.

(2) Smaller systems will become available with high quality output for restricted tasks and domains, especially in combination with controlled languages supported by appropriate authoring tools.

(3) Small systems capable of generating multi-lingual text or speech output on the basis of tabular or other types of structured input (as opposed to free text input).

(4) Multilingual information systems based on spoken dialogues.

Typical examples:
- Translators work benches;
- Meteo (TAUM, Canada), weather report translation (the classical example);
- Patrans (CST, Denmark), patents translation; - Restricted English for translation and documentation purposes (Bull, Perkins Engines, Xerox);
- Multilingual traffic information via car radios (Bosch, Philips).

4. Trends in research

Strange enough, the trends in research as reflected by scientific journals such as Machine Translation and conferences such as TMI (Theoretical and Methodological Issues in MT) seem to have a fairly loose relationship with what is happening on the market.

As a matter of fact, where market developments tend to go towards more specialized and restricted systems, the general trend in research seems to be to widen the traditionally rather narrow horizon of the MT research community. Three main dimensions can be identified:

(1) From sentence to text: Discourse-based translation, essential for, e.g., pronominal reference problems.

(2) Looking beyond linguistics: Domain-based translation, where disambiguation takes place on the basis of domain restrictions.

(3) Looking beyond traditional methods: Combinations of traditional (rule-based) and statistical approaches, where preferred linguistic analyses and translations are chosen on the basis of statistical data, gathered from large mono- and bilingual corpora.

5. Conclusions for the TELRI project

Current trends in MT, both in research and on the market, show that there are a number of areas where a massive effort with respect to linguistic resources is needed:

- First of all, there will be an increasing demand for lexical and terminological resources in all Eastern European languages in order to serve as the basis for better tools for professional translators.
- Secondly, the definition of appropriate controlled or subdomain languages will only be possible on the basis of large collections of real-lifa data.
- Thirdly, the increasing popularity of statistical methods in MT will require the availability of more and more mono- and bilingual text corpora.

6. For more information

Hutchins and Somers. 1992. "An Introduction to Machine Translation". Academic Press, London; (encyclopaedic overview).

Arnold et al. 1994. Machine Translation, An Introductory Guide, Blackwell, Manchester; (introduction).

Machine Translation (editor Sergei Nirenburg). Kluwer, Dordrecht; (scientific journal).

Language Industry Monitor (editor Colin Brace). LIM, Amsterdam; (journal, with excellent market information).

European Association for Machine Translation. ISSCO. Geneva; (professional organisation of users and researchers).

MULTEXT-EAST: Multilingual Text Tools and Corpora for Central and Eastern European Languages

Tomaž Erjavec*, Nancy Ide**, Vladimír Petkevič***, Jean Véronis**

* Laboratory for Language and Speech Technologies
Institute Jožef Stefan
Jamova 39
61111 Ljubljana
Slovenia
Tel.: +386 61 1773 ext. 507
Fax: +386 61 219 385
E-mail: tomaz.erjavec@ijs.si

** Laboratoire Parole et Langage
Centre National de la Recherche Scientifique et Université de Provence
29, Av. Robert Schuman
F-13621 Aix-en-Provence Cedex 1
France
Tel.: +33 42 204 356
Fax: +33 42 205 905
E-mail: veronis@cs.vassar.edu

*** Institute of Theoretical and Computational Linguistics
Faculty of Philosophy, Charles University
Celetná 13
110 00 Praha 1
Czech Republic
E-mail: vladimir.petkevic@ff.cuni.cz

BEST COPY AVAILABLE

1. Introduction

The language industries rely increasingly heavily on the availability of large-scale language resources, appropriate software tools, and standards to make them maximally reusable. Such resources and tools exist or are under development for most western languages, and efforts to develop standards for corpus encoding and linguistic software development are well underway, in particular in the LRE project MULTEXT, one of the largest EU projects in the domain of language tools and resources (Ide and Véronis, 1994).

However, there have been no comparable efforts for Central and Eastern European (CEE) languages. No large-scale, systematic attempts at corpus collection currently exist (in particular for multilingual, parallel corpora in these languages); tools specifically adapted to corpora in CEE languages are not widely available; and most standardization efforts have not yet taken into account the specific characteristics of CEE languages.

MULTEXT-EAST is a spin-off of the LRE project MULTEXT which is intended to fill these gaps by developing significant resources for six CEE languages (Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovenian) and by adapting existing tools and standards to them. MULTEXT-EAST extends MULTEXT's scope to CEE languages with the following goals:

- test and adaptation of language standards
- development of an annotated multilingual corpus
- development of morpho-lexical resources
- adaptation of the MULTEXT corpus tools.

MULTEXT-EAST began at approximately MULTEXT's mid-point, at a time when MULTEXT's specifications, methods and tools were well-developed enough to extend to additional languages. At the same time, it has been possible to incorporate feedback from application to vastly different language types (especially Slavic and Finno-Ugric) while specifications, methods, and tools are still under development.

Together, MULTEXT and MULTEXT-EAST create a unique network of more than 20 academic research centers and companies, all developing and using common lingware and methodologies for 13 EU and CEE languages. Moreover, MULTEXT-East will also coordinate its efforts in tool adaptation with the TELRI concerted action, esp. with Working Group for Tool Availability. This working group will promote the MULTEXT tools and help in adapting them to the MULTEXT-East languages and different software platforms.

2. Corpus

2.1 Markup

MULTEXT has developed a Corpus Encoding Standard (CES) (Ide and Véronis, 1995b) optimally suited for use in corpus linguistics and language engineering applications, which can serve as a widely accepted set of encoding standards for European corpus work. The standard identifies a minimal encoding level that corpora must achieve to be considered standardized in terms of descriptive representation (marking of structural and linguistic information) as well as general architecture (so as to be maximally suited for use in a text database). It also provides encoding conventions for more extensive encoding of linguistic corpora and for linguistic annotation.

The CES is an application of SGML (ISO 8879:1986, Information Processing-Text and Office Systems--Standard Generalized Markup Language). It is based on and in broad agreement with the TEI Guidelines for Electronic Text Encoding and Interchange (Sperberg-McQueen and Burnard, 1994; see also Ide and Véronis, 1995a). The TEI Guidelines were expressly designed to be applicable across a broad range of applications and disciplines; therefore, they treat not only a vast array of textual phenomena, but are also designed with an eye toward the maximum of generality and flexibility. The CES, on the other hand, treats a specific domain and set of applications, and can, therefore, be more restrictive and prescriptive in its specifications. In addition, because the TEI is not complete, there are some areas of importance for corpus encoding that the TEI Guidelines do not cover. Therefore, the first major task in developing the CES has involved evaluating, adapting, selecting from, and extending the TEI Guidelines to meet the specific needs of corpus-based work.

In its present form, the CES provides the following:

- a set of metalanguage level recommendations (particular profile of SGML use, character sets, etc.);
- tagsets and a DTD for documentation of the encoded data;
- tagsets, DTDs, and recommendations for encoding textual data, including written texts across all genres, for the purposes of corpus-based work in language engineering.
- tagsets, DTDs, and recommendations for encoding linguistic annotation, including segmentation, grammatical annotation, and parallel text alignment.

MULTEXT-EAST is applying the CES to texts in six CEE languages, including fiction and newspaper data. The experience of applying the CES to

these new languages has led to a major revision and extenstion of the CES, in particular to handle the required additional character sets. In addition, the lack of substantial pre-existing texts in some electronic format in the Eastern European countries and the resulting need to develop many corpora based on printed materials only has made it necessary to consider the kinds of markup that can or should be included and the optimal stages of markup enhancement when corpora are generated in this way.

2.2 Corpus composition

MULTEXT-EAST is building an annotated multilingual corpus, composed of material comparable to MULTEXT, whose primary goal is to provide an example and test-bed for:

- the applicability of MULTEXT's multilingual tools (especially enginebased tools, alignment software, and multilingual extraction tools) to CEE language corpora;
  and
- the applicability to CEE languages of the TEI Guidelines and MULTEXT's TEI-based corpus markup standard, as well as the MULTEXT-EAGLES pan-European lexical specifications and part-of-speech tagset.

The sample corpus is being prepared in TEI-conformant SGML format and annotated for basic structural features as well as sub-paragraph segmentation, part of speech, and alignment of parallel texts.
The sample corpus will be composed of three major parts:

(1) Multilingual Comparable Corpus
For each of the six MULTEXT-EAST languages, the comparable corpus will include two subsets of at least 100 000 words each, consisting of
- fiction, comprising a single novel or excerpts from several novels;
- newspapers.

The data will be comparable across the six languages in terms of the number and size of texts. Selection criteria will be applied to each subset to ensure quality. The entire multilingual comparable corpus is being prepared in CES format, manually or using ad-hoc tools, and will be automatically annotated for tokenization, sentence boundaries, and part of speech annotation using the project tools. For each language, a portion of the corpus will be hand validated.

87

(2) Multilingual Parallel Corpus
For the six MULTEXT-EAST languages, the parallel corpus will include approximately 100 000 words per language, consisting of translations of Orwell's Nineteen Eighty-Four. The entire multilingual parallel corpus will be prepared in CES conformant format, manually or using ad-hoc tools, and then automatically annotated using the project tools. For each language, half of the corpus will be marked and validated for alignment and sentence boundaries. Alignment will be between the English version and each of the six MULTEXT-EAST languages, thus, constituting six pair-wise alignments. A portion of the corpus will be hand validated.

(3) Multilingual Speech Corpus
MULTEXT-EAST will record a small corpus of spoken texts in each of the six languages, similar to the EUROM-1 speech corpus, comprised of 40 short passages of five thematically connected sentences, each spoken by several native speakers with phonemic and orthographic transcriptions. MULTEXT-EAST will enhance this spoken corpus with markup for prosody, segmentation, and part of speech. The prosody markup will consist of two levels: F0 curve modeling and symbolic coding. This markup will be performed using the tools developed in MULTEXT, and a portion of the corpus will be hand validated. The orthographic transcriptions will be marked for tokenization, sentence boundaries, and part of speech annotation, and they will be hand validated. The project will carry out a restricted alignment, consisting of the alignment of word boundaries as well as the beginning of accented vowels between signal and transcription for one speaker per language.

3. Morpho-lexical resources

An important aspect of tool development in MULTEXT is the engine-based approach, where all language-dependent materials (lexicons, morphological rules, etc.) are provided as data. MULTEXT-EAST, in collaboration with EAGLES, has evaluated, adapted, and extended the specifications (rule format, lexical specifications, corpus tagset, etc.) for the language-dependent material developed in MULTEXT to cover the six MULTEXT-EAST languages (Monachini, 1995). Accomodating the different language families represented among the MULTEXT-EAST languages has demanded substantial assessment and modification of the pre-existing specifications, which were developed for Western European languages only. The work carried out in MULTEXT-EAST has, thus, broadened the base and

contributed significantly to defining a universal mechanism for lexical specification.

MULTEXT-EAST is developing the following language-specific resources for use with the various annotation tools:

(1) Segmentation rules. This includes rules describing the form of sentence boundaries, quotations, numbers, punctuation, capitalization, etc.

(2) Special tokens. The language-specific data required by the segmenter includes lists of special tokens (frequent abbreviations and names, titles, patterns for proper names, etc.) with their types.

(3) Morphological rules. The project is providing morphological rules for the MULTEXT-EAST languages, which are needed by the morphological tools. The rules provide exhaustive treatment of inflection and minimal derivation. Each lemma in the lexical lists used by the project (see below) is associated with its part(s) of speech and morphological rules.

(4) Lexical lists. For each of the six MULTEXT-EAST languages, a lexical list containing at least 15 000 lemmas is being developed for use with the morphological analyser. Each entry includes the following information: inflected-form / part of speech / morphological information / lemma. A mapping from the morpho-syntactic information contained in the lexicon to a set of corpus tags (used by the part of speech disambiguator) is also provided according to the MULTEXT tagging model (Véronis and Khouri, 1995).

4. Tools

4.1 Standardization

There is a serious lack of generally usable tools to manipulate and analyze the text and speech corpora and collections that are now becoming widely available. The linguistic software that exists at present only begins to cover growing needs. Industrial software is often expensive or unavailable, and is usually hard to adapt or extend. On the other hand, the substantial body of natural language processing academic software is often experimental, hard to get, hard to install, under-documented, and sometimes unreliable. In both cases, tools are typically embedded in large, non-adaptable systems which are fundamentally incompatible. Worse, there is enormous duplication of effort: it is not at all uncommon for researchers to develop tailor-made systems that replicate much of the functionality of other systems and in turn create programs that cannot be re-used by others and so on in an endless software waste

cycle. Although efforts to develop standards for data representation are underway, little effort has been made to develop standards for linguistic software, and software reusability is virtually non-existent.

MULTEXT has joined efforts with the EAGLES sub-group on Tools to address this need by working towards the establishment of Guidelines for Linguistic Software Development (LSD) (Véronis and Ide, 1995). These guidelines specify a general lingware development environment, including recommended standards for all aspects of software development, data representation, linguistic annotation, etc. The establishment of such a set of guidelines enables the interchange of tools and data among researchers and sites, compatibility among tools with potentially diverse functionality, and in general contributes to the creation of reliable, high quality tools.

Standards exist or are being developed in many areas relevant to linguistic software development, including

- character sets
- document encoding
- language and country codes
- application program interfaces
- programming languages
- internationalization and localization of programs
- etc.

Each of these standards covers a small piece of what would serve as a general lingware development environment, but none has been developed with an eye toward the overall coherence of such an environment. The goal of the MULTEXT/EAGLES LSD Guidelines is to bring together existing or emerging de jure or de facto standards sufficient to address the scope of an entire Linguistic Software Development system.

MULTEXT tools are intended to demonstrate many of the basic principles of software development that will be recommended in this environment, including especially atomicity and language-independence. MULTEXT-EAST provides a significant test-bed for the MULTEXT tools, in particular because these principles are aimed towards enabling easy modification and extension to new (and possibly very different) languages.

4.2 Adaptation of Multext tools

MULTEXT is developing a set of corpus manipulation tools that is freely available, coherent, extensible, and language-independent, including:

Morphosyntactic tagging:
- segmenter: marks sentences, quotations, words, abbreviations, names, etc.;
- lexical lookup and morphological analyser: provides lemmas, morphological features, and parts of speech;
- part-of-speech disambiguator: disambiguates parts of speech where alternatives exist;

Parallel text alignement:
- aligner: provides alignments of sentences among parallel texts;

Prosody tagging:
- signal editor and signal analysis utilities (MES)
- prosody tagger (MOMEL): derives automatic modelling of F0 curve and symbolic coding of intonation from the speech signal;

Corpus manipulation tools:
- SGML query language (SgmlQL);
- format conversion utilities;
- multilingual string manipulation library;
- post-editing tools: assist in hand validation of automatically annotated corpora.

The tools are implemented under UNIX. All MULTEXT tools are designed using an engine-based approach where all language-dependent materials are provided as data. Therefore, extension of the tools to cover CEE languages in MULTEXT-EAST primarily involves providing the appropriate tables and rules for these languages. However, some adaptation of the tools is expected, given the potential for new problems which may be posed by these vastly different language types (i.e., languages with heavy inflection, free word order, etc.).

5. Conclusion

MULTEXT-EAST is extending the MULTEXT effort to six CEE languages by adapting MULTEXT's tools, developing linguistic resources for these six languages, and providing a multilingual corpus comparable to the

one developed for EU languages within MULTEXT. This will validate and enhance MULTEXT's tools and its software and markup standards. Most importantly, it will enable not only early use of developing standards in CEE countries but also the possibility for feedback as a result of adaptation to a vastly different set of languages.

As in MULTEXT, all of the work within MULTEXT-EAST will be performed in conjunction with EAGLES and the TEI, and thus provide an extension and validation of the work of these initiatives on standardization to a new range of languages. Similarly, like MULTEXT, MULTEXT-EAST will distribute its results, tools, and corpora and linguistic resources for six CEE languages free or at cost by ftp and CD-ROM.

References

Ide, N., and J. Véronis. 1994. "MULTEXT (Multilingual Tools and Corpora)". Proceedings of the 14th International Conference on Computational Linguistics, COLING'94, Kyoto, Japan 1994, 90-96.

Ide, N. and J. Véronis (eds.). 1995a. The Text Encoding Initiative: background and context. Dordrecht: Kluwer Academic Publishers.

Ide, N. and J. Véronis. 1995b. Corpus Encoding Standard. Document MUL/EAG CES1. <URL:http://www.lpl.univ-aix.fr/projects/multext/CES/CES1.html>

Monachini, M. (ed.). 1995. Common Specifications and Notation for Lexicon Encoding of Eastern Languages.

Deliverable 1.1. Multext-East Project COP-106. ftp:////www.lpl.univ-aix.fr/pub/multext/docs/ME1.1.tex

Sperberg-McQueen, C.M. and L. Burnard. 1994. Guidelines for Electronic Text Encoding and Interchange. Chicago and Oxford: Text Encoding Initiative.

Véronis, J. and N. Ide. 1995. Guidelines for Linguistic Software Development. Document MUL/EAG LSD2. <URL:http://www.lpl.univ-aix.fr/projects/multext/LSD/LSD2.html>

Véronis, J. and Khouri. 1995. Etiquetage grammatical: mod+le. <URL:http://www.lpl.univ-aix.fr/projects/multext/LEX/LEX2.html>

Appendix: Project's fact sheet

MULTEXT-EAST Participants

PART	PARTICIPANT'S FULL NAME	CC R
AIX	Laboratoire Parole et Langage Centre National de la Recherche Scientifique	FR C
PISA	Istituto di Linguistica Computazionale Consiglio Nazionale delle Ricerche	IT A
SOFIA	Department of Mathematical Linguistics Institute of Mathematics Bulgarian Academy of Sciences Sofia (Bulgaria)	BU P
PRAG	Institute of Theoretical and Computational Linguistics Charles University Prague (Czech Republic)	CZ P
BYLL	BYLL Software, Ltd. Prague (Czech Republic)	CZ S
TARTU	Laboratory of the Estonian Language Tartu University Tartu (Estonia)	EE P
BUDA	Linguistic Research Institute Hungarian Academy of Sciences Budapest (Hungary)	HU P
MORPH	MorphoLogics Budapest (Hungary)	HU S
BUCHA	Research Institute for Informatics Bucharest (Romania)	RO P
ICI	ICI Bucharest (Romania)	RO S
LJUBL	Laboratory for Language and Speech Technologies Institute "Jožef Stefan" Ljubljana (Slovenia)	SI P
AMEB	AMEBIS Ljubljana (Slovenia)	SI S

Abbreviations:

PART: Participant's short name
CC: Country Code
R: Role (C- Coordinator, P- Full partner,
A- Associate partner, S- Subcontractor)

Effort: 345 person-months
Duration: 24 months
Start state: 1 May 1995

Contact point:
Dr. Jean Véronis (coordinator)
Laboratoire Parole et Langage
CNRS & Universite de Provence
29, Av. Robert Schuman
13621 Aix-en-Provence Cedex 1 (France)
Tel.: +33 42 95 36 34
Fax : +33 42 59 50.96
E-mail : veronis@univ-aix.fr

Speech Recognition
A General Overview

Luis de Sopeña

IBM S. A.
Madrid Scientific Centre
Santa Hortensia 26–28
E-280 002 Madrid
Tel.: +341 397 5752
Fax: +341 519 3990
E-mail: Lsopena@vnet.ibm.com

1. Areas of Speech Processing

There are five main areas in the field of Speech Processing:
1) *Speech Coding* deals with the compression of the digital representation of the speech signal in order to facilitate economical transmission or storage.
2) In *Speech Synthesis*, a synthetic speech signal is created from preexisting text with an attempt at reaching maximum intelligibility and naturalness.
3) Using techniques for *Speaker Identification*, the machine identifies the speaker by his/her voice in order to ensure restricted access to information, computer, or the physical premises.
4) In *Speech Recognition*, the information in a spoken message is identified so as to have the computer perform the corresponding command or transcribe in written form the dictated text.
5) Finally, *Spoken Language Translation* deals with two-way communication via speech: a spoken message is identified, translated into a different language and this translation synthesised in speech form, in order, e.g., to enable a dialogue between speakers of different languages.

2. Difficulties in Speech Recognition

There are some well-known difficulties in the field of speech recognition, shown in the list below:

- The variability of sounds (words, phrases, subword units), within a single speaker and across different speakers.
- The variability of channel, depending on the characteristics of the different types of microphones.
- The variability of background noise: side conversations, street noise, telephone rings, etc.
- The variability of speech production, which adds spurious sounds to words proper (mouth clicks, hesitations, breath noise.)

3. Main Functions of Speech Recognition

Speech recognition can be used in a variety of situations:
1) To perform *Query* operations, such as the consultation via telephone of a bank for account balances, the consultation of phone information lines for theatre schedules and the like, and also for phone call transfers.

2) *Data entry* situations may include the giving of a credit card number, dialing from mobile phones, filling out forms, and booking airline reservations.
3) *Command and Control* operations in which speech recognition is important occur when the hands and/or eyes are busy, during menu navigation and machine control, and while completing dark room work.
4) Speech recognition plays a key role in *dictation* when entering free text into a computer via speech.

4. Technical Characteristics of Speech Recognition Systems

The technical characteristics of speech recognition systems depend on several variables, the most important of which are the following:
1) The *vocabulary size* can range from small (10-100 words) for simple commands, to medium (1000 words) for form filling, or to large (more than 20 000) for such complex situations as dictation.
2) Other than vocabulary size, the *speaker dependence* of a given system can vary from being trained to a specific speaker, to being adaptive to each user as (s)he speaks, or even speaker independent.
3) The *speaking mode* varies between continuous text and isolated words, where pauses between words are needed for an adequate recognition.
4) Speech recognition systems can be *domain dependent*, meaning they can only recognize a constrained syntax (e.g., a list of commands or of questions), or independent, where free text can be dictated.
5) *Multiple language support* is also an important characteristic.

5. Knowledge Sources in Speech Recognition

The knowledge sources in speech recognition are based on three different models:
1) *Set of Phoneme Models:*
Reference to the typical sound of a phoneme, specified by the probability distribution of its spectral and temporal properties.
2) *Word Lexicon:*
Represented as a sequence of the above phonemes (Acoustic Model).
3) *Language Model:*
Statistical model extracted from large corpora of texts.

6. Speech Recognition Process

The objective of the speech recognition process is to determine the sequence of words which caused most probably the observed sequence of acoustic vectors (see figure 1).

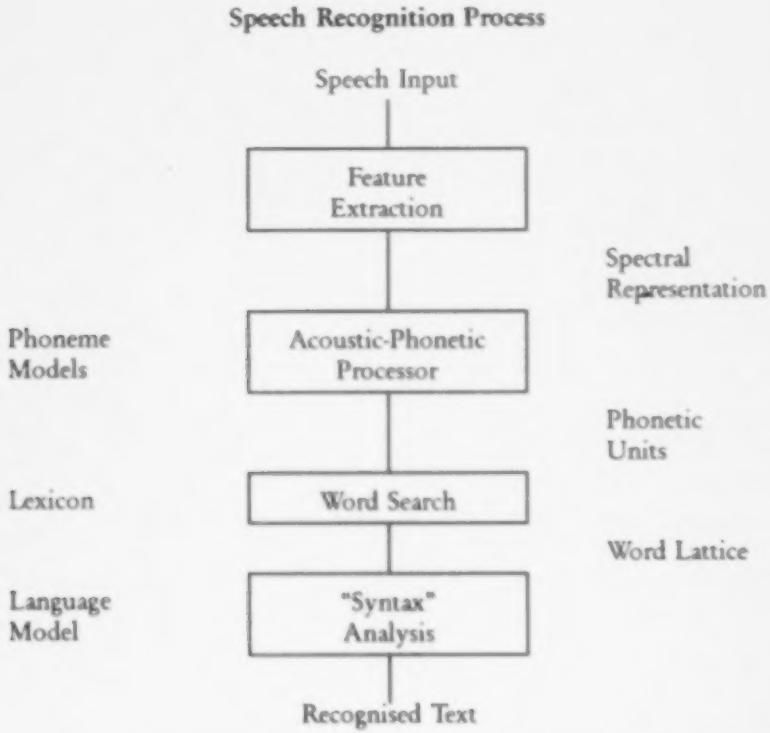


Figure 1. Determine sequence of words which caused MOST PROBABLY the observed sequence of acoustic vectors

7. Speech Recognition Today and Tomorrow

An example of a present-day Speech Recognition system is the *IBM VoiceType Dictation System*. Its most important characteristics are:
- Works on a 486 SX 25
- Recognises more than 30K different words
- Needs a short enrollment process
- Recognises discrete speech (with small pauses between words)
- Able to handle 70–100 words per minute
- Available for 6 languages
- With a very high recognition rate (>96%)

Tomorrow, however, research is promising much more. Speech recognition systems will be able to handle:
- Any speaker, without need for training
- Continuous speech
- Very large vocabularies (more than 250K words)
- With telephone capabilities
- Including natural language understanding
- On Personal Digital Assistants

These systems will be used in dictation; phone mail, DB access, home shopping, translation, and much more.

Most important of all, Speech will be an "enabler", i.e., existing and new applications will be accessible using speech.

8. Main Players in the Field of Speech Recognition

The main players in the field of speech recognition are the following:
- the European Community
- ARPA (Wall Street Journal Contest, Air Travel Info Service (ATIS))
- Industrial Research (IBM in dictation, AT & T for phone services, and many smaller companies)

One of the continual points of discussion in the field of speech recognition is the relative importance of English as compared to other languages. But nonetheless, speech systems are developed for other major languages as well (e.g., French, German, Spanish).

Language Resources: The Foundations of a Pan-European Information Society

Wolfgang Teubert

Institut für deutsche Sprache
Postfach 1016 21
D-68016 Mannheim
Tel/Fax: +49 621 1581 415
E-mail: wolfgang.teubert@ids-mannheim.de

0. Language resources for language technology: academia meets industry

Language engineering is the core of information technology, and information technology will be the key industry of the 21st century. The information super highways conceived today will soon transport infinite amounts of digital data, images, sounds, tables, figures, calculations, and process protocols. If these data are to be intelligible, if they are to make sense, they must be bound together by language. Without natural language processing (NLP), information remains incomprehensible.

More than any other continent, Europe is multilingual by common commitment. This situation provides a challenge to European language technology. We all want information to cross borders freely; however, countries can only uphold their cultural and linguistic identity if all the relevant information is available and accessible in the national language(s). This is an important principle of the European Union today, and it also holds for all European nations who have not yet joined the European Community. For the emergent European information society, we have to develop a language technology that meets the multilingual challenge. It will have to support the production, revision, conversion, presentation, publication, documentation, and last, but not least, translation of texts in technical and everyday language; and it will have to grant language-independent retrieval by sophisticated interaction modes based on natural languages. This demand calls for a modular solution. Language independent algorithms have to be combined with language specific software. The goal must be to impose as few restraints as possible on one's language in computer-aided communication activities. A "controlled documentation language" designed to avoid all kinds of ambiguities and undesirable complexity will impede the generation and transmission of novel content and concepts. A language without fuzziness, opacity and metaphorisation brings about a petrification of ideas. Only principally unrestricted use of one's language (and the reunification of any "controlled documentation language" in man-machine interaction will guarantee the creativity of information exchange. If this challenge is met, European language engineering will play a leading role on the global market.

The quality of all language technology rests on the linguistic knowledge determining the algorithms of any NLP application. This linguistic knowledge is accessible in and from language resources. We find it in scientifically designed text corpora and in corpus-based or, at least, corpus-validated lexicons; we can extract it from textual and lexical resources by recursive generalisation processes and convert it into the form needed in a specific application. For corpus analysis and lexicon generation, we need powerful generic soft-

ware that will have to be modular to take care of language specific and language independent aspects. It would be erroneous to believe that the linguistic knowledge needed for sophisticated language technology applications is already available. So far, language studies have indeed accumulated huge repositories of data and their interpretation. However, these data were collected for human users only. Human users are able to draw on analogies and can apply inductive reasoning. They have common sense and understand (to some extent) the world they live in. Therefore a human translator cannot be a model for machine translation, nor can bilingual dictionaries be the source for lexicons in translation systems. The task of machine translation (MT) can be compared to the task of translating a text on a remote subject field from an unknown language A into an unknown language B with the help of a bilingual A/B dictionary. A MT lexicon would have to contain all the information to make such a translation possible. The necessary linguistic knowledge has to be gathered from scratch. We do not find it in existing dictionaries. There is no alternative to text corpora.

Corpora are the raw material of all language technology. The better they are, the more expensive is their creation. Language industry, small-and medium-sized enterprises in particular, often cannot afford to build them up. On the other hand, in practically all European countries there are focal language centers with a long tradition in the creation (and also in the application) of language resources. This is not only true for all of Western Europe, but also for most of Central and Eastern Europe including the former Soviet Union. In its short history, computational linguistics always has been a global discipline; and the NLP community was and continues to be well connected.

However, while the results of academic research traveled freely (with a few clandestine pockets here and there), language resources, corpora, and machine readable dictionaries did not flow as easily. Due to the lack of hardware compatibility, the old restrictions now often have given way to new limitations based on property rights. Solutions only can be found in joint efforts of research and industry. We need a European network of all academic research institutions to allow the fee flow of language resources between all partners under fair conditions.

A research project, however, is only the first step. In a second step, we have to set up an operational infrastructure of (public domain) research and (private) industry. We need a common platform where providers and users of language resources come together, share expertise, discuss their needs, exchange resources, join forces, and give birth to new visions. Private industry will ensure that new language technology applications find a market (and the money invested will not be wasted), and public domain research will provide the linguistic expertise to make the products a success.

In some European countries, such a national infrastructure already exists; in others, it is gradually evolving. Still most of the work is devoted to monolingual applications. Until some years ago, there was not much cross-border cooperation at least not in academic circles. This is why repeated efforts in Western Europe have been made to set up a transnational infrastructure that can serve the needs of multilingual language technology applications. In March 1995, the European Language Resources Association (ELRA) was founded with strong backing by the Commission of the European Community. It will be seen in the years to come which additional measures are necessary to secure Europe's role as a leading actor in the global language engineering market. This future Europe will be larger than the European Union of today. Linguistic expertise, language resources, and computational linguistics are highly developed in the countries of the former Soviet Union and in Central and Eastern Europe. We can observe the emergence of a dynamic, if still small, language industry in these locations. If we want to make Europe a competitor on the world market of language industry, we must build up a common infrastructure all the way from Galway at the West Coast of Ireland to Vladivostok at the Sea of Japan.

In most European countries, we find one or more focal academic institutions dedicated to document, analyze, and describe the national language(s) and to put linguistic knowledge across to the public. Language was regarded as the paramount cultural asset responsible for ethnic or national identity. In all research work, the main emphasis was, therefore, on this language; it was important to demonstrate that this language was as well-researched as any other major world language. Communication across borders, from language to language, if encouraged at all, was rarely a topic for these institutions. The transnational flow found its limitations in the number of translators and translations, serving as a useful filter of incoming and outgoing information.

Today, the situation has changed: European integration is now imperative to our societies. A precondition is the uncurtailed accessibility of information. National economies can successfully compete on the global market only if all relevant data are available. However, although it is relatively easy to obtain all the information wanted, it is usually written in a foreign language, unless you happen to speak English or perhaps French as your mother tongue. So, the large majority of people who do not speak a foreign language are put at a disadvantage. German companies regularly complain that calls for tender issued by the Commission of the European Communities are initially and at times only published in French and English: this makes it much more difficult for them to compete. International TV channels first thought that they could reach their audience by transmitting their programmes in English;

however, programme sponsors now insist on subtitles or dubbing in the local languages. Most viewers prefer programmes in their native language even if they understand English.

Language, once a cultural asset, has now become an economic commodity. The focal national language institutes have acquired new responsibilities. Now, their task is to provide the means that information from abroad can be made available in the national language and that locally produced information can be distributed world-wide in major international languages and in the language of the neighboring countries. To train and employ more translators is not sufficient. We have also to take care that the necessary language technology is being developed.

1. Current Issues in Language Engineering

Spelling checkers were among the earliest successful language technology applications. They have been accepted as useful devices and are still being sold today in ever-improved versions. On the other hand, more ambitious projects often have failed. The majority of early machine translation systems, particularly the more sophisticated ones have not survived. SYSTRAN is still kept alive by the European Commission's translation services, but many others have disappeared without leaving a trace.

Spelling checkers do not need semantics. Even in the seventies, they were based on little more than a list of the most frequent word forms, i.e., linguistic knowledge widely available or easy enough to generate from corpora. Machine translation, however, needs semantics. Our understanding of word meanings or lexical semantics in the seventies was contained in dictionaries, and it was arranged in a form that a human user, with some experience, could understand, using a great deal of implicit knowledge about the world and inductive reasoning and having the ability to draw analogies–all assets computers usually do not possess. Therefore, it is no surprise that early applications involving semantics were not very successful.

Today, we know that the semantic data needed for sophisticated language technology cannot (or only to a very small extent) be derived from dictionaries designed for human users. Rather, it has to be generated from scratch, namely, on the basis of corpora. Language technology can work with two kinds of semantic information. The one is rule-based and presupposes an intellectual semantic analysis of the phenomenon in question. The other kind uses statistics and is not really semantic at all: it computes the five or ten words or word forms preceding or following the word in question and relates

this information to the different translation equivalents found in parallel corpora. The German word *Schnecke*, e.g., is translated into English either by *snail* or *slug*, where *snail* refers to the creature with a 'house' and *slug* without one. The rule-based approach states just that and searches the German text for clues from which we can infer the correct translation equivalent. The statistic-based approach does not look at the meaning at all. It looks for words and other traces that frequently occur when *Schnecke* is translated as *snail* and for other patterns co-occurring with *Schnecke* being translated as *slug*. In the context of *slug*, we would probably find words like vegetable *(garden), lane, wet*, and various forms of *get rid of*; while in the case of *snail*, I would expect words like *table, course, wine*, but also *vineyard* and *sunny*.

Today's successful applications involving semantics work with an amalgam of the rule- and the statistic-based approach. The statistical approach has some attractive advantages: the data required can be generated form corpora with no or only little human intervention; and since it is just an emulation of semantics, one does not have to be able to state explicitly what a lexical item means. Indeed, it leaves out the entire question of meaning. Its inherent shortcoming is the rate of accuracy. Even a rate of 95% of correct translation equivalents implies that every twentieth word in a text is mistranslated, practically every sentence and certainly more than what most people would like to live with. On the other hand, the rule-based approach can be a very expensive alternative. It presupposes something like a bilingual dictionary that would enable translation of a text correctly into an unknown foreign language. The reason why such dictionaries do not exist for human users or machines is that the explicit linguistic knowledge they would have to contain is not yet available and that this knowledge is be extremely expensive to produce.

The new generation of language technology applications (monolingual and bilingual or multilingual ones) deals with semantic problems. They recognize the fact that computers cannot understand spoken or written texts in the way humans can. Therefore, these text processing systems can only emulate the human faculty of 'understanding' by a mixture of rules and probabilities. To find the right mix is less a question of theory and principles than of calibration and learning by doing. The crucial point is the performance of an application under real life conditions. The application has to prove its cost efficiency, i.e., it must demonstrate that it can complete a task cheaper than a trained human. After two decades of experimental and pilot systems, the emphasis today is on robust applications for which there is a real market, i.e., one where users are willing to pay a fair price.

This new state of the art is reflected in the policy of the European Commission on information technology. Under the Telematics 4th Framework

Programme, only such projects are funded where the usefulness of the application under development is verified and attested by an industrial interest group, representing the end users of the application. In addition, the project consortium must include partners from academic research and industry. Industry can claim only 50% of its development cost for funding. Thus, the companies involved in such a project must hope to gain the other 50% of their financial investment from product sales. Market orientation is a top priority for the Commission. This is reflected in the list of topics for funding mentioned in the Work Programme of the 1994-1998 Telematics Applications Programme. (1) I will present the following topics:

- Document Creation and Management
- Information and Communication Services
- Translation and Foreign Language Acquisition
- Globalization and Localization of Software

1.1 Document Creation and Management

The Work Programme lists the following objectives: *to produce high-quality documents faster and more economically; more effective handling of and localization of product documentation; faster and more accurate drafting in a non-native language; improved checking of style, grammar and comprehensibility; more consistent use of technical language and terminology; better reuse of existing texts; transparent access to in-house and remote document translation services; automatic indexing and categorization of documents; ability to locate and browse documents; content analysis; automated generation of documents from database records.* (2)

The following applications are mentioned in the Work Programme:

- Authoring of office documents
- Collaborative authoring of technical documents
- Document management
- Document interchange and message handling
- Report generation

All of the applications make use of tools that have existed for some time now such as bilingual lexicons, thesauri, terminological databases, text editors, content analysis procedures, knowledge extraction processes etc. Some of them deal only with formal aspects of texts like editors including spelling and syntax checkers; some use rule-based semantics like (monolingual or multilingual) thesauri, particularly in connection with word sense disambiguators. Content analysis and automatic indexing is based on statistic methods. The

applications mentioned bind together a number of tools in order to be able to fulfill a particular task, a task that would otherwise have to be carried out by humans. The goal is to relieve people from tedious routine like repeating ready-made text elements and to give them more time for writing new text segments.

These procedures also can ensure completeness and consistency of highly formalized texts. Minutes of meetings are a typical example for a document type that is strictly formalized in composition and that involves much routine. As has been shown, adequate software can improve the quality of employment references by asking the author to go through exhaustive check lists.

The quality of applications in this field depends on the architecture that holds together the modules. This relates to the linguistic and technical specifications for the data interchange between the various tools contained in an application. The other important quality parameter is the fine-tuning or calibration of the system for the particular task for which it was designed. What is an appropriate relationship between stylistic variation and consistency in the use of terminology? How much choice should the author have in the organization of the document, in the vocabulary, or in the complexity of syntax he is to use? Less freedom will be permitted if the documents are to be machine rather than hand translated.

1.2 Information and Communication Services

The Work Programme lists the following objectives: *better and more selective data access; more intuitive language interfaces; content-based information analysis and filtering; technology-mediated interpersonal communication services with advanced spoken and written language capabilities; content-based information browsing and navigation; maximum localization of human-computer interfaces.* (3)

The following applications are mentioned in the Work Programme:

- Text- and voice-based communication services
- Access to information and transaction services
- Mobile information and communication services
- Interface and software globalization

As we can see from the objectives mentioned, emphasis in this topic is on semantics: 'more intuitive language interfaces', 'content-based information analysis', 'content-based information browsing'. For applications in this field, it will not be sufficient to use existing tools and assemble them as modules in

a new configuration. Of course, there are already many speech recognition systems. While the majority of them are still experimental, some are already being marketed to end users. But they are still far away from being satisfactory. Speech to text software, dictation systems still impose severe restrictions on the user, e.g., making distinct pauses between the words. They need extensive training for each speaker, and their accuracy rate makes post-editing necessary. Better performance can be expected from speech understanding systems that will look for the most plausible reading of an utterance and not just match the phonetic input with a possible orthographic output. But, most of those systems are still in their pilot stages. To make their performance more reliable, we will need more semantic data that can tell us which reading is more plausible than the alternative ones.

More semantic data than currently available are also needed for any kind of content-based information extraction. As I said before, these data should not be taken from existing dictionaries but from corpus analysis. Content-based analysis talks about concepts and the semantic relationship between concepts, not about expressions and the syntactic relationship between expressions. The quality of information extraction rests with the thesaurus and its links of synonymy, hyponymy etc. that it establishes among lexical units. Most existing dictionaries show these links only between single words, but this view is too narrow. One of the results of the computational analysis of language is that the lexical units to be considered often are multiword units, collocations, and phraseologisms. Applications in this field will have to use better lexicons describing such entities if they want to be superior to existing tools. Building corpus-based lexicons, however, is rather expensive. Indeed most existing lexicons have been converted from dictionaries which are not corpus-based. The creation of a new generation of standardized homogeneous generic corpus-based lexicons for the languages of Europe is a major task, and it certainly costs more than small and medium-sized enterprises can afford. This job also demands the kind of linguistic expertise one finds more in academic institutions. Close cooperation between academic research and private industry, therefore, is an important condition for the development of robust applications in the information and communication services.

1.3 Translation and Foreign Language Acquisition

The Work Programme lists the following objectives: *lowering of language barriers; more effective support for professional and occasional translators; access to in-house and remote translation facilities; better support for the translation of*

technical documents; speech translation of business documents and correspondence; more effective learning of lexical and grammatical knowledge; more effective foreign language reading and writing aids. (4)

The following applications are listed in the Work Programme:

- Translator's toolbox and translation aids
- Machine-aided translation
- Tutored and self-trained foreign language acquisition

It is noticeable that the goal of machine translation per se has been given up. The huge EUROTRA project has failed in a number of aspects; machine translation of unrestricted text without extensive post-editing is not possible. Still, machine translation is better than no translation--it gives the reader some idea of what the text is about. Machine translation is possible and cost-effective (in spite of the necessary post-editing) where huge amounts of relatively standardized texts in a specific domain (i.e., airplane maintenance manuals) have to be translated. Its quality can be improved by restricting the input: restrictions on the vocabulary and the syntactic complexity of sentences, on the use of anaphoric pronouns, on coordination structures etc. Thus, it is no surprise that weather forecast messages are a typical example for successful automatic translations from French into English and vice versa.

In most situations, however, it is preferable to develop tools enabling human translators to produce better translations in less time. In this field, again, semantics is the core issue. The meaning of a text should remain constant, regardless of the language used. The task of creating bilingual or multilingual lexicons is enormous. Existing dictionaries are of little help. They are good enough for human translators to render a foreign language text into their native language. However, due to a lack of relevant semantic data, they are not sufficient to translate from one's native language into a foreign language.

The basis of new tools to facilitate human translations should include parallel corpora, i.e., texts and their translations into other languages. They contain, though only in an implicit form, all the linguistic knowledge a translator needs; however, this knowledge has to be extracted from the corpus of parallel texts before it can be reassembled in tools that are the core of novel translation aids. Here again, the lexical unit that is translated is not only the single word, but (more often) multiword units, collocations, and phraseologisms. Bilingual and multilingual lexicons will be repositories of translation equivalents, regardless of whether they are single words or complex phrases extracted from parallel corpora and arranged according to linguist principles. Therefore, the methodology for lexicons of translation equivalents will probably be developed and perfected in academic research. On

the other hand, how much of the information and in which sequence it will be presented to the translator is a matter of fine-tuning and will vary from application to application.

The conception of new translation aids, such as the translation memory, must have a high priority in European language technology. For a long time, academic linguistic research and the variation of language resources have focused on the monolingual situation. However, if information is to travel freely in a multilingual Europe, language barriers have to be lowered considerably by new translation tools. Multinational software houses were the first to take up the challenge of multilingual language technology. Yet, even they lack the specific language resources needed for a robust performance of multilingual applications. Their creation is the most immediate task for all the nations desiring participation in the European information society.

1.4 Globalization and Localization of Software

The field of globalization and localization of software is not a separate chapter in the Work Programme of the new Telematics Applications Programme. However, it is briefly mentioned under the heading 'Information and Communication Services': Multilingual products and services presuppose flexible and effective methods and tools for human-computer interface and *software internationalization and localization, leading to faster and more widespread user acceptance.* (5)

From the point of view of academic institutions and small- and medium-sized private industry in Central and Easter European countries and of the former Soviet Union, the more interesting aspect is globalization of resources, generic tools, and specific applications. There are linguistic data, and there is sophisticated software, but it is not used outside the country or language area where it was designed. Of course, there are language specific tools like lemmatizers and morphological generators for which there seems to be no use outside the particular language; however, a tool like a lexicographer's workbench or, to be even more general, a tool for language data compression can be used for any language. Complex applications usually will be conceived in a modular way. Some modules will be language specific, while others are language independent. The emergent European language resources and language technology infrastructure will provide a platform where academic and industrial partners in numerous countries can be found willing and competent to adapt existing tools to other languages and to distribute such tools in Europe and on a global scale.

This is true not only for tools but also for corpora and lexical resources. Even if they were originally composed for monolingual applications only, it is worthwhile turning them into multilingual resources. Some texts in a corpus might have been translated in other languages; if these translations are included, then there is a new parallel corpus for which there will be demand in other parts of the world as well. There are several ways to establish semantic links between entries of monolingual lexicons. Thus, it is possible to constitute a multilingual lexical database on the basis of existing monolingual lexicons. These lexical data can then be used as input for the alignment of translational equivalents in parallel texts, and this procedure in turn supplies new semantic links between the entries of different languages. Globalization of lexical resources is a precondition for the development of all multilingual applications and should be given high priority. The evolving European infrastructure must encourage joint activities in this field.

Localization of software that has been developed for the English-speaking market is the other face of the coin. Localization is necessary for complex tools designed for nonspecialist end-users, whenever the language barrier would keep them from using it. This is clearly the case for software like text editors, style checkers, or other authoring tools for any software that handles or manipulates language. Every language specific module has to be redesigned for the local language; also, some adjustments usually have to be made for the more language independent modules. For these and for all other tools, which are to be used by a wider audience who does not necessarily understand English, texts used in front ends have to be localized, and the user manuals have to translated. There is a quickly growing demand for localization services; and for academic institutions and small local software houses, it can be very attractive from the financial point of view to offer this kind of service. Contacts with the huge global and predominantly English-speaking market can, thus, be established and used later for other activities as well.

Information technology is constantly undergoing change, change that is reflected in the terminology of this discipline. Institutions working in the localization field have immediate access to new terms as they occur in manuals and other technical product descriptions. Thus, these institutions have a strong voice in the creation of the national terminology. Experience has shown that whoever can exert control over the terminology of a given discipline also often has a leading edge over his or her competitors. This is another reason why it can be attractive to set up a localization service.

One of the three 'colleges' in the field of language resources (besides speech and written language) is terminology. Key industries today are organized as global markets, and their terminologies are strongly dominated by the English

language. In order to sell products on local markets, however, advertisements, product descriptions, maintenance manuals, and other technical documents have to be available in the language(s) of the respective country. In the seventies and the early eighties, some international companies believed that people would buy their products even if they were described only in English; however, there is a lot of evidence that whenever consumers had an alternative they preferred the product that was fully documented in their native language. The localization of international (mostly English) terminology is becoming an important activity in the field of language resources. The European Commission has funded terminology infrastructure projects in Western Europe and also, under the COPERNICUS programme within all of Europe including the Commonwealth of Independent States. (6)

2. The Importance of Language Resources

As we have seen in the preceding chapter, the quality of language technology applications rests foremost with the comprehensiveness and reliability of the language data with which the tools are used. For traditional applications like a spelling checker, 'hard' data are needed: word forms (including orthographic variants) and a morphosyntactic analysis leading to lemmatization. For more sophisticated applications that somehow involve meaning, we need 'soft' data, data which are interpretations of hard data by competent linguists. Finally, for multilingual applications, we need soft data for various languages and procedures allowing the mapping of these data onto each other, again necessitating interpretation by competent linguists.

Hard and soft data together constitute our linguistic knowledge. We have argued that today's knowledge is neither reliable enough nor sufficient for the development of robust heavy duty tools performing noticeably better than the existing toy or pilot systems.

Many computer linguists, particularly those with an engineering background, take it for granted that, with the linguistic knowledge available today, a steady improvement of language technology systems is possible to the point where, say, an operational, robust system for machine translation can be developed. They are impressed by the fact that seemingly astonishing results can be achieved with stochastic methods necessitating little knowledge and even less interpretation of soft data of the language(s) involved. They believe to be justified in their convictions because they look at natural languages as nothing but a set of particularly complex formal languages. However, there is a generic difference. Given basic conditions, formal languages

can be translated into each other. But anyone who has translated from one natural language to another knows it takes more than just the hard linguistic data plus a few statistic operations. Translators must also understand the content of the text and must be aware that they will understand it only if they have sufficient knowledge of the world. This is not knowledge about facts rather now we interpret the objective world that we perceive. To some extent, this (community agreed) interpretation will have to be integrated into the language data.

As I stated before, the linguistic knowledge available--the knowledge formulated in existing grammars and dictionaries--is unsuitable for language technology tools for two reasons. First, most of it is not corpus-based; rather it reflects the individual linguist's (lexicographer's) competence based on a collection of data (citations); and, however large their collection may be, it is permeated by a bias that cannot be avoided. Second, traditional grammars and dictionaries have been devised for human users, and human users differ substantially from machines. Humans use inductive reasoning and can draw analogies easily; faculties like these are taken for granted and are reflected in the traditional arrangement and presentation of processed linguistic data. Language technology tools cannot take recourse to common sense. for language technology applications, all knowledge has to be spelled out in the form of rules, lists, and probabilities.

This task is sufficiently demanding in itself. In order to carry it out, we have to go back to the sources; and the only source for linguistic data is the corpus, the authentic and actual texts in their unannotated representation. But, anyone who has gone to the sources has also experienced the problem that when we start analyzing language as it occurs in a corpus, we gain evidence that renders existing grammars and dictionaries as very unreliable repositories of linguistic knowledge. We discover that our traditional linguistic knowledge gives us a very biased view of language, a view that has its roots in the contingency of over two thousand years of linguistic theorizing. We are so accustomed to this view that we take it for the truth, for reality, not just for an interpretation of hard data. It is true that traditional grammars and dictionaries have helped us, fairly satisfactorily, to overcome the linguistic problems we humans have to deal with. But they will not suffice for language technology applications.

This is why, however cumbersome and expensive it may be, language has to be described in a way that will be appropriate for language engineering. In the Council of Europe corpus project - the *Multilingual Dictionary Experiment* (project leader, John Sinclair, with participants from Croatia, England, Germany, Hungary, Italy, Sweden) - it has been demonstrated that

monolingual and bilingual dictionaries are of no (or only little) use when it comes to automatically translating a word from one language into another in cases where there is more than one alternative. (7) A close analysis of the problems involved in the translation of nominalizations between German, French, and Hungarian (also corpus-based) has shown that all the descriptions available in traditional dictionaries and grammars are inadequate, incomplete, and ultimately useless. (8) To reduce the cost of a corpus-based language analysis from scratch, which is indispensable, corpus exploitation tools have to be developed which arrange the hard facts (including statistic-driven devices for context analysis) and which process them (with a great deal of human intervention for the semantic interpretation of data) into algorithmic linguistic knowledge, into rules derived from objective data rather than individual competence. Perhaps, this will result in the finding that traditional categories like noun, verb, and adjective do not, after all, reflect categories useful for NLP.

Corpora are the basic language resources. They have been used by linguists for about thirty years. We have learned that the early discussion on the representativeness of corpora led in the wrong direction. Corpora represent nothing but the texts they consist of, and certainly not a language universe. The analysis of corpora has given rise to new insights, particularly concerning the vocabulary. In the sixties, we were accustomed to think that we could define the core of general language as the intersection of special languages and that such a definition would make it possible to define a finite general language vocabulary of perhaps 50 000 to 100 000 lemmata. Today, we are much less sure about the usefulness of such constructs. Instead, we talk about balanced corpora, corpora composed according to parameters like text type (e.g., 3rd person, present tense only), genre (e.g., romance, instruction, free asymmetrical conversation) and domain (e.g., angling, crime, stock exchange).

We know that a balanced corpus regardless of size, number of parameters, and numbers of values assigned to each parameter does not represent general language; instead, we say it can be used for a number of different purposes. It does not represent the vocabulary of general language because a general language vocabulary is not a meaningful concept. All we can say is that we aim at a corpus that is 'saturated' in terms of the vocabulary. This means that a particular chunk of our balanced corpus representing one text type, one genre, and one domain (e.g., texts in the 1st and 3rd person, past, and present tense; newspaper diary; angling) is saturated once the growth rate of the vocabulary stops decreasing and becomes constant. There is no point from which there will be no hitherto unrecorded words, but there is a point from which there will be perhaps eight new words (types) for each 10 000 ad-

BEST COPY AVAILABLE

ditional words of text (tokens). Saturation of corpora is a fairly new concept, and no one knows yet what it leads to in terms of corpus size.

John Sinclair has recently developed a corpus typology, and I am using it in my brief account of corpus types: (9)

Special corpus. Special corpora are assembled for a specific purpose, and they vary in size and composition according to their purpose. By intention, special corpora are not balanced (except within the scope of their given purpose) and, if used for other purposes, give a distorted view of the language segment. They can have a number of advantages compared with balanced corpora. Their main advantage is that the texts can be selected in such a way that the phenomena one is looking for occur much more frequently in special corpora than in a balanced corpus. A corpus that is enriched in such a way can be much smaller (perhaps ten times) than a balanced corpus providing the same data.

Reference corpus. Reference corpora come closest to the old concept of a representative corpus. They are composed on the basis of relevant parameters agreed upon by the linguistic community and should include spoken and written, formal and informal language representing various social and situational strata. The idea behind reference corpora is that they can be used for a large variety of purposes, thus, rendering most special corpora unnecessary. They are also the point of reference when it comes to measuring the distortion of special corpora. They are used as benchmarks for lexicons and for the performance of generic tools and specific language technology applications. They are large in size; 50 million words is considered to be the absolute minimum; 100 million will become the European standard in a few years.

Monitor corpus. Language changes, and these changes should be reflected in a constant growth rate of corpora, leaving untouched the relative weight of its components (i.e., the balance) as defined by the parameters. The same composition schema should be followed year by year, the basis being a reference corpus with texts spoken or written in one single year.

Opportunistic corpus. The opportunistic corpus is an inexpensive alternative to the reference corpus. It is a collection of electronic texts that can be obtained, converted, and used free or at a very modest price; and its composition principle is that one should take all one can get and try to fill in blank spots as soon as they are recognized. Their place is in environments where size and corpus access do not pose a problem. The opportunistic corpus is a virtual corpus in the sense that the selection of an actual corpus (from the opportunistic corpus) is up to the needs of a particular project. Today's monitor corpora usually are opportunistic corpora.

Comparable corpus. For multilingual research and applications, corpora in each language are needed that follow the same composition pattern and, thus, can be used for language comparison. Opportunistic corpora cannot fulfill this claim. The focus is, therefore, on reference corpora. The Commission of the European Community is funding a project whose main goal is the creation of comparable reference corpora (of 50 million words each) for all the official languages of the European Union including Catalan and Irish. Comparable corpora are an indispensable source for bilingual and multilingual lexicons and a new generation of dictionaries. (10)

Parallel corpus. Texts in one language and their translations into other languages constitute parallel corpora. They are the source for the detection of translation equivalents and, thus, can play an important role in the development of multilingual lexicons. In order to do this, parallel corpora must be aligned at least sentence by sentence, preferably phrase by phrase. Their disadvantage is that the language of translations is distorted and does not contain the full range of vocabulary and syntax. To compensate for this deficiency, one can set up reciprocate parallel corpora, corpora containing authentic texts as well as translations in each of the languages involved. This allows double-checking translation equivalents. Only if a collocation or phraseologism occurs also in authentic texts, is it counted as an acceptable equivalent.

3. Standardization and Validation

Anyone who wants to build up a corpus has to make a number of decisions about how to encode the text. A text is more than a stream of words; it contains much more information. It may have one or more authors, a title, chapter headings, tables and figures, footnotes, foreign language citations, font shifts, paragraphs, and many other features. What should be retained? How should it be encoded? Should one aim at reconstruction of the source text complete with its specific layout? There are no strict rules. However we decide, we have to abide by the decision we made for the entire corpus. If we mark footnotes in text A, we should also mark them in exactly the same way in text B. There must be an explicit list of the features to be encoded, and there must be a set of unambiguous codes. Only a text collection that has been encoded in such a consistent way should be called a corpus.

Why do we need consistency in the creation of corpora? Corpora are raw data waiting for linguistic analysis. To extract information from corpora, we have to use software. The more complicated our query is, the more sophisti-

cated our software must be. If we want to extract complex phraseologisms consisting of all headlines of newspaper texts containing a certain phraseologism, we must make certain that headlines are marked consistently.

That corpora must be marked up consistently is self-evident. All large centers for language resources have developed or obtained corpus exploitation software demanding explicit coding of such linguistic and extra linguistic features. Within the limitations imposed by the software, the centers were free to chose the codes. This was adequate for a time when there was little exchange of corpora and software. However, when there was more demand for language resources and when a new generation of corpus software had to be developed able to handle corpora of 100 million words or more and still yield results in interactive access in a reasonable time, cooperation, exchange, and distribution became more and more important.

In recent years, the Text Encoding Initiative (TEI), an international project with strong North American and Western European participation, has developed standards and guidelines for the encoding of all sorts of texts (spoken and written) as corpora to be used as language resources. Likewise, standards and guidelines were developed for the set-up and exchange of lexical and terminological data. (11) The TEI recommendations are based on the SGML standard of the International Standards Organization (ISO). In the present form, they demand much extra work for the corpus compiler if they are to be followed step by step--probably more than most institutions are willing to invest. Therefore, European language resources projects like the Network of European Reference Corpora (NERC) and the first PAROLE project singled out subsets of the TEI standards that should be adopted by any new corpus (and lexicon) project. (12) ISO is preparing a number of SGML-based standards for terminology databases and the exchange of terminological data. (13)

For some years now the European Expert Group on Language Engineering Standards (EAGLES), an activity funded by the European Commission, have developed recommendations for the linguistic (and to some extent also extralinguistic) annotation of corpora and the data categories used for lexicons. (14) For some application areas, there are only preliminary reports; in other fields, operational guidelines are already available.

Building up an infrastructure of language resources only makes sense if these resources are standardized. As we said before, this is the only way to ensure the reusability of resources for unlimited applications within the global NLP community. Standardization also helps in making resources comparable and in building links between resources. Only if monolingual lexicons follow the same architecture and use the same categories, is it possible to merge them into a multilingual lexicon. All alignment software for parallel corpora oper-

ates on the assumption that all relevant phenomena are encoded consistently. Standardization is a precondition of an operational language resources infrastructure. Its importance cannot be exaggerated.

On the other hand, it is not necessary to adopt these standards for internal use in institutions. Whoever has assembled a consistently encoded corpus and has developed proprietary software for corpus exploitation is under no obligation to change it to the TEI or EAGLES standards. These common international standards have to be used when it comes to data interchange. If we want to exchange or distribute our resources or our tools, we must convert them to be conformant or at least compatible with the narrow set of recommendations agreed upon by the NLP community. This task can be carried out by appropriate conversion software, some of which is offered by companies or is available under public domain.

What exactly should be standardized? It is a hot issue of discussion how far standardization should stretch. Basically there are two fields where standards and guidelines are offered for corpora. The one concerns the encoding of all relevant features of the text that we want to include into a corpus, data that are there, e.g., in the printed text of a novel. We usually can identify the author(s), find the titles, determine if there are chapters, paragraphs, and other layout features serving a purpose for the text. Features like these have to be encoded. But, what about full stops? These dots at the end of a sentence have to be preserved; however, do they also have to be disambiguated? On the surface, they do not differ from the little dot at the end of an abbreviation, e.g., viz. or etc. There are reasons to disambiguate them: we have to know where a sentence starts and ends. If we want to search for the co-occurrence of specific words forming a phraseologism it is useful to confine the search to sentences as linguistic units. If we want to align parallel text, the minimal parameter for alignment is the sentence, marked by the full stop and not to be confused with the little dot indicating that the previous string is an abbreviation. But if we decide for disambiguation, we add linguistic information to the text, using our linguistic knowledge and competence. Any kind of corpus annotation means adding even more and potentially more questionable linguistic information. Admittedly, there is a lot of agreement among experts whether a given word is a noun, verb, adjective, or something else. Yet, how should we tag the first element of language technology, the English equivalent to the German word Sprachtechnologie? Is it really a noun, or a noun used as a modifier, like an adjective, or is it just the first constituent of a compound?

If we add linguistic information of this kind to a corpus, then we might obliterate the very objective we wanted to achieve by using the corpus. Cor-

BEST COPY AVAILABLE

pora are used for obtaining new linguistic knowledge. But if we use traditional knowledge when it comes to the extraction of evidence from a corpus, we may never get to the point where we may want to discard old categories and replace them by new ones confirmed by corpus evidence. It is from looking at unannotated corpora that we have learned a lot about the fuzziness of the concept of lexical units and that we now see a continuum from a word segment like -euro- via single words, multiword units, collocations, to whole phaseologisms. Particularly in multilingual applications, it is obvious that the translation equivalents we are looking for are only very rarely single words and sometimes can be complete phrases or even sentences.

Any linguistic information added to a text or corpus tends to replicate any bias inherent in that category. Therefore, we must be careful that standardization of linguistic features originally not present in a text but added by hand or automatically does not obscure the evidence we expect to obtain from corpora. Normative so-called language-independent tagsets and categories for word class, morphosyntactic or even semantic information, whether in corpora or in lexicons, are useful only insofar as the can establish comparability between resources of different origin. If these tagsets are used for extracting evidence for linguistic phenomena from corpora, results have to be evaluated with great care.

The creation of language resources is expensive. The more effort that has been spent on the composition, documentation, and encoding of textual features on the annotation of added linguistic information, and on the comparability with related resources, the more valuable a corpus (or a lexicon) will be. To use it for just one specific purpose would be a waste--it should be reused as often as possible, covering a wide range of applications. As we said above, standardization is an absolutely necessary precondition if the resources are not only to be used internally within one single institution but are also to be made available to the NLP community. While in the past the exchange of resources and resource-oriented tools were based on bilateral agreement between provider and user, distribution centers today serve as platforms for the language resources market. They will be accepted only if the material they offer for distribution proves valuable to the clients, and therefore, it has to be fully documented and standardized. The first center was the Linguistic Data Consortium (LDC) at Pennsylvania University, which was created more than five years ago. Recently, the European Language Resources Association (ELRA) was founded for the (Western) European market. It will be complemented by national language resource centers. The Trans-European Language Resources Infrastructure (TELRI), a Concerted Action funded by the European Commission, is serving as a bridge between Central and Eastern Europe

(including the Commonwealth of Independent States) on the one side and Western Europe on the other. These distribution centers will pass on submitted resources to academic and industrial users and will collect license fees. It is no longer necessary to contact the institution where the resources were created in first place.

Such an arrangement works only if the resources are not only documented but validated as well. What does validation mean for written resources? For corpora, it means that the corpus has the size it claims, that it is composed the way it claims, that it is encoded the way it claims, that all features encoded can be used for retrieval, that annotations used conform to a given standard, and, above all, that the error rate for encoding and annotation does not exceed a certain level. Validation guarantees the client that he gets what he ordered and that he can rely on the resources to the extent stated by the validation certificate.

Validation has to be carried out on an unbiased and neutral basis, and this means not by the institution where the resources were created. Some of the validation features are language-independent, like the conformancy of the text representation with accepted rules and standards for encoding. But, some are language-specific, e.g., a claim that a lexicon really covers the core vocabulary of a given language, that the meanings assigned to a lemma have been gained from corpus evidence, and that the error rate of morphosyntactic tagging does not exceed a certain level. Therefore, it makes sense that validation for written resources should be carried out by institutions where there is sufficient competence for the language in question. On the other hand, to make validation reliable, all recognized validation centers should follow a validation procedure prescribed by a standard that is accepted by the NLP community. Some projects for the development of validation procedures are under preparation; the European Commission has included validation as a new work item in the new Telematics Programme.

Validation is important for generic resources designed for reusability. These resources will always be necessary for language technology. However, the more sophisticated applications are being designed, the more specific the language resources backing them will have to be. These resources will either have to be produced from scratch or by adding value to existing generic resources. In either case, specific resources of this kind will probably not be distributed by large centers, but produced by one institution and then be passed on directly to the user on a bilateral agreement. This trend will lead to a reassessment of the validation issue in the long run.

4. The Trans-European Language Resources Infrastructure (TELRI)

The boom in language industry has brought with it a growing demand for more and better monolingual and multilingual resources. In Europe, only a joint effort of existing focal language institutions could be expected to harmonize existing and designed standardized new resources in compliance with the needs of dictionary makers and developers of language technology applications.

The European Commission, realizing the central role of language engineering in the emergent information and communication technology market, has supported a number of relevant infrastructure activities, serving the needs of the three 'colleges' – speech (spoken language), terminology, and written resources. These projects helped to set up a common infrastructure for the countries of the European Union and the European Economic Area (formerly EFTA), and at the same time encouraged formation of national language resources networks. After years of preparation, the new PAROLE II project with partners from all European Union countries will produce a first generation of harmonized, comparable generic textual and lexical reusable resources, meeting the basic demands of language technology.

But Europe is larger than the European Union. All European countries must be given the opportunity to participate on an equal level in academic and industrial research and development. In the COPERNICUS Programme, the European Commission provided a framework of projects aiming at the integration of activities in Central and Eastern Europe with complementary ones in Western Europe. Several projects currently underway deal with various aspects of speech, terminology, and written resources. One of these projects dealing primarily with written resources is the Trains-European Language Resources Infrastructure (TELRI). (15) Since it aims at including as many partners in Central and Eastern Europe as possible, it is set up as a Concerted Action rather than as a project proper. Its partners are 22 focal language and language technology institutions in 17 countries, six Western European and 11 Central and Eastern European countries. The partners in the West are also cooperating in the PAROLE projects, thus linking closely TELRI activities with Western European developments. For the time being, there are no partners from former Yugoslavia (with the exception of Slovenia) or from the Commonwealth of Independent States. However, formalized links have been established with the leading institutions in Croatia, Serbia, and Russia; and these associated partners are participating in TELRI activities.

The Concerted Action TELRI has an initial duration of three years, beginning in early 1995, and is working on a budget of about half a million ECU.

It is not a research project: rather, its goal is to create a viable infrastructure in order to establish a permanent platform for industry, research institutes and universities, and to supply the NLP community with precompetitive or public domain monolingual and multilingual language resources. These resources are: corpora, machine readable dictionaries and lexicons, lexical data bases, and software tools for the creation, reuse, maintenance, valorization, and exploitation of linguistic data.

The activities of TELRI are organized in Working Groups for specific tasks. The collection, documentation, and dissemination of relevant information on language resources, providers and users, their potentials, and their needs is a basic activity. TELRI will promote the formation of national language resource networks, and TELRI partners will act as focal nodes. They will also design small scale joint ventures with private industry in order to foster cooperation between academic research and development. TELRI will pool and enhance existing service activities, providing resources, expertise, consulting and training facilities. The central platform will be annual seminars directed at the needs of small- and mediu[illegible]sized enterprises. TELRI will engage in European and global standardization and validation activities and contribute to the harmonization of already existing resources. It is organizing joint research in the field of corpus-based multilingual lexicography and the use parallel aligned texts.

The main impact of TELRI will be to make the whole of Europe a strong competitor in the emerging market of language engineering and communication technology. Central and Eastern European language centers will participate in the creation of harmonized monolingual and multilingual resources making them ideal partners for joint ventures. Their huge potential in the creation of language data and the development of generic and specific language-related software will help to establish Europe as a leading force in language technology.

References

1. European Commission - DG XIII (ed). 1994. Telematics Applications Programme (1994–1998). Work Programme 15 December 1994. Luxembourg: Publications of the European Commission (1994), pp. i-x (further referred to as: EC - DG XIII (1994))
2. EC - DG XIII (1994), Appendix A5, p. 2
3. EC - DG XIII (1994), Appendix A5, p. 4
4. EC - DG XIII (1994), Appendix A5, p. 5

5. EC - DG XIII (1994), Appendix A5, p. 4 (my italics)
6. European Commission - DGs XII, III, XIII, XVII (ed). 1995. INCO COPERNICUS Information Package. Edition 1995/1996. Luxembourg: Publications of the European Commission (1995).
7. Teubert, W. 1996. "Parallel Corpora and Multilingual Lexicography". In: International Journal of Lexicography. (Forthcoming)
8. Teubert, W. 1992. "Zur Behandlung von Präpositionalattributen im Wörterbuch". In: Cahiers d'Études Germaniques. Beiträge zur Lexikologie und Lexikographie des Deutschen 23/1992, p. 119–135
9. Sinclair, J. and J. Ball, 1995. "Text Typology" (External Criteria). Draft Version. Electronic Document on the Pisa EAGLES ftp server, Birmingham
10. LE-PAROLE Consortium (ed). 1995. "LE-PAROLE Annex I" (Technical and Financial Annex). Annex to the project proposal submitted under the 1994-1998 TELEMATICS Applications Programme of the European Commission. Pisa/Paris: Consorzio Pisa Ricerche/GSI-Erli (1995), Chapter 3.1.2
11. Inquiries to: C. Michael Sperberg-McQueen, University of Illinois at Chicago, Academic Computing Center (M/C 135), 1940 W. Taylor, Rm. 124, Chicago, IL 600612-7352, USA, E-mail: u35395@uicvm.uic.edu
12. NERC Consortium (ed). 1993. Network of European Reference Corpora. Final Report. Birmingham/Leyden/Malaga/Mannheim/Paris/Pisa (1993); PAROLE (final report to be published in spring 1996; inquiries to Professor Antonio Zampolli, ILC Pisa, E-mail: eagles@ilc.pi.cnr.it
13. Inquiries to: Secretariat of ISO TC 37, c/o DIN Berlin, Burggrafenstraße 6, D-10787 Berlin, Germany
14. Information on EAGLES guidelines and recommendations can be obtained through the EAGLES ftp server at Pisa: nicolet.ilc.pi.cnr.it (anonymous login)
15. Information via WWW: http://www.ids-mannheim.de/telri/telri.html or E-mail: telri@ids-mannheim.de

Rail-lex Slovenia – A Modern Railway Dictionary

(Joint Venture Case Study)

Primož Jakopin

Institute for the Slovenian Language ZRC SAZU
Novi Trg 4
SLO-61000 Ljubljana
Tel.: +386 61 1256 068
Fax: +386 61 1255 226-253
E-mail: primoz.jakopin@uni-lj.si

BEST COPY AVAILABLE

The two partners involved in the joint venture are the Railway Traffic Institute, part of the Slovenian Railways and the Institute for Slovenian Language, which is part of the Scientific Research Centre at the Slovenian Academy of Sciences and Arts. The work on the project, Dictionary of Railway Terminology (Železniški terminološki slovar), began in January 1994 and is to be completed in five years by the end of 1998.

The dictionary is part of a larger European undertaking, Rail-lex Europe, which is a coordinated effort of 29 members of the UIC, Union Internationale des Chemins de fer (International Union of Railways). UIC consists of 97 railway and transport organizations from Europe and other parts of the world. The aim of the Rail-lex project, which started in 1990 and produced a CD-ROM Rail Lexic with 12 000 headwords and 102 topics in 4 languages (English, German, French, and Italian) in 1994, is to put together a modern, multilingual communication infrastructure. This infrastructure should promote links between the railways themselves, between railways and industry, and between research and commerce and should contribute to the standardization of railway terminology.

Table 1 gives the main topics of the UIC dictionary (in German to illustrate better the engineering spirit of the field):

Maschinentechnik
- Kraftwagen, Schiffe
- Fahrdynamik, Spurfuehrungstechnik
- Fahrzeug Eisenbahn-Triebfahrzeug (gem. Terme)
- E-Trieb-fahrzeug
- Brennkraft-Triebfahrzeug
- Dampf-Triebfahrzeug
- Wagen (gemeinsame Terme)
- Wagen: Personenverkehr
- Wagen: Gueterverkehr
- Fahrzeugteil, -aufbau (gemeinsame Terme)
- Fahrzeugteil, -aufbau Personenverkehr
- Fahrzeugteil, -aufbau Gueterverkehr
- Kupplung
- Laufwerk
- Wagenkastensteuerung
- Fahrzeugantrieb Kraftuebertragung
- Bremse, Bremsdienst
- Elektrotech. Anlagen, Elektr. Energie

Werke
Nachrichtentechnik
Personal, Soziales
Recht
Planung, Kontrolle, Revision
Rechnungswesen
Unternehmenspraesentation
Organisation, Datenverarbeitung, Information
- Managementorganisation
- Bueroorganisation und -kommunikation
- Datenverarbeitung
- Information, Dokumentation, Statistik

Finanzen
Verkehrswesen, Marketing, Angebotsentwicklung
Verkauf, Verkaufsabwicklung, Servicebetriebe
Produktionbetr. Infrastrukturplanung
Allgemeine Technik, Neue Verkehrssysteme
Bautechnik, Neubaustrecken

Table 1. The main topics of Rail Lexic, in German

Two of the main topics, the largest and the one related to data processing, are shown with their respective subtopics as well.

Rail-lex is coordinated by UICs European Rail Research Institute (ERRI), which is based in the Netherlands. The languages of the partners who have participated from the beginning of the project are English, German, French, Italian, Spanish, Esperanto, Hungarian, Dutch, Polish, Portuguese, and Swedish. In the course of the project, the language pool has been enlarged by Bulgarian, Croatian, Czech, Danish, Lithuanian, Norwegian, Persian, Romanian, Russian, Slovak, and Slovenian.

On the Slovenian side, the head of the project is mag. Peter Verlič, leader of the team at the Railway Traffic Institute, who is aided by Marjan Vrabl and his team of railway experts in Maribor, the second largest Slovenian city. That is also where, at the same time, a new set of railway codebooks, manuals, and other documentation are being prepared. Since Slovenia gained independence in 1991, the changes needed to bring the railway closer to UIC standards have to be made. The bulk of the keywords from Rail Lexic have now been translated; and together with additional keywords, which reflect the social and other specific circumstances in Slovenian railway, they now form the first draft of a 15 000-headword dictionary:

čas prometne konice	peak-traffic hours peak-traffic period rush hours rush period
direktni vozni list : vozni list na celotni prevozni poti	through consignment note
drobljenje = delitev blokov informacij na fragmente zahtevane velikosti za prenos (maks. dolzina paketov je določena z omre žjem), ki se posredujejo naenkrat	fragmentation into packets = Division of information blocks into fragments of the size required for transmission (maximum length of packets is defined by the network) which will be routed individually
duplikat voznega lista	duplicate of the consignment note

Table 2. Headword samples

The draft of the dictionary will be open to criticism from the railway itself and from the wider audience until the end of 1996 when a linguistic revision on the side of the Institute for the Slovenian Language will also be completed.

A New Dutch Spelling Guide

J.G. Kruyt, P. G. J. van Sterkenburg

Institute for Dutch Lexicology INL
P.O. Box 9515
NL-2300 RA Leiden
Tel.: +31 71 527 2270
Fax: +31 71 527 2115
E-mail: kruyt@rulxho.leidenuniv.nl

128

1. Introduction

The *Institute for Dutch Lexicology (INL)* is a research institute subsidized by the Dutch and Belgian governments. Apart from corpus-based lexicography, INL is active in the field of the compilation and (semi-)automatic linguistic annotation of text corpora, and the development of retrieval systems which allow the user to consult the corpora along various parameters. Corpus development at INL dates from the mid-seventies. Up to 1990, the INL text corpora were developed for lexicographical purposes mainly. Presently, they are used for a broad variety of research and applications (cf. Kruyt 1995, Van Sterkenburg & Kruyt 1996).

The language data available at INL have proven to be of interest for external users as well. The majority concerns academic users. The more limited interest in the data from the part of commercial companies may be explained by investments being relatively large and the market restricted, due to the rather small Dutch speaking language area and a minor interest of computational linguists in the Dutch language. Still, INL is presently cooperating with two commercial publishing houses in the framework of product development. A product recently developed in cooperation with *AND Electronic Publishing* (Rotterdam, The Netherlands) is a CD-ROM containing the historical dictionary of the Dutch language *Woordenboek der Nederlandsche Taal (WNT)*. The present paper reports on the development of two corpus-based Dutch spelling guides, in which the INL data have been crucial for both the list of entries and the values (actual word forms) for the information categories per entry. Partners were the publishing house *Staats Drukkerij en Uitgeverij (SDU)* (The Hague, The Netherlands) and, as for the second one, the Dutch-Belgian government body *Nederlandse Taalunie (NTU)* as well. The spelling guides are characterized in Section 2, the relevant INL data in Section 3. The final section discusses the need for more harmonization in the development of and access to data, so as to make product development more efficient and feasible.

Rather than on proper spelling issues, this paper focuses on the conditions in which the spelling products were developed. For detailed information on Dutch spelling, we refer to Molewijk (1992) and De Vries, Willemyns & Burger (1993).

2. Dutch Spelling Guides

The two corpus-based spelling guides compiled by INL date from 1990 and 1995, respectively. The guide of 1990 includes an earlier one, published in

1954. The latter was preceded by the first spelling guide commonly accepted, published in 1866. A concise characterization of the two earlier guides serves as a background for the products of 1990 and 1995.

2.1 Spelling Guide 1866

The *Word List for the Spelling of the Dutch Language* (1866) was compiled by the founders of the historical dictionary of the Dutch language *"Woordenboek der Nederlandsche Taal" (WNT)*, M. de Vries and L.A. te Winkel. In the framework of dictionary design, they devised a spelling system based on etymological principles. In 1882, the Dutch government implicitly acknowledged this spelling system as the official one, by applying it in the criminal code. A slightly simplified version of the spelling system got an explicitly official status in the next century, in 1947.

From the outset, the spelling system was criticized. It was considered too complicated, particularly for the children at primary school, being obliged to correctly apply the spelling system.

2.2 Spelling Guide 1954

In order of the governments of the Netherlands and Belgium, a new spelling guide was compiled by a committee of twelve Dutch and Belgian (Flemish) experts in the field. They developed a more simplified spelling system. The compilers additionally had to bridge the gap between the Dutch and Flemish views on spelling issues. The Flemish traditionally wished to distinguish themselves from the French, and hence, had a preference for the character "k" over "c", e.g., "publikatie" rather than "publicatie" ("publication"). The Dutch on the other hand, did not appreciate German-like spellings at the time (a few years after the Second World War), and they preferred "c" over "k": "publicatie". The problem mainly applies to loan words. As a compromise, the *Word list of the Dutch language*, published in 1954, often lists two spelling forms for a word, a "preferred" one and an "allowed" one.

The double spelling forms evoked much criticism and the spelling was still considered too complicated. Additionally, many new words came into use, which were not in the list. The governments of The Netherlands and Belgium felt responsible for a solution and installed many spelling committees. For political reasons, a decision remained forthcoming until 1994.

BEST COPY AVAILABLE

2.3 Spelling Guide 1990

Meeting the need for an orthography for many words that entered the Dutch language since the fifties, INL and the *Staats Drukkerij en Uitgeverij (SDU)* published an inofficial spelling guide in 1990, the *Revised word list of the Dutch language*. SDU is a privatized, commercial publishing house. The activities include the publication of books (language, art, history) and state products (passports, bank notes etc.), as well as database publishing. INL was responsible for the contents of the guide, SDU for its publication. The division of the revenues was established by contract. The guide was published in printed form only; a slightly encoded machine-readable version was available for internal use.

The guide of 1990 contains the word list of 1954 (ca. 65 000 entries), and additionally ca. 30 000 new entries. The list contains entries in the "preferred" spelling (2.2.) only; an appendix lists entries in "preferred" and "allowed" orthography. The information categories per entry (microstructure) include the entry (in "preferred" orthography), variant form (if relevant), hyphenation, indication of meaning (in case of homographs), genus for nouns, inflected forms (plural and diminutive for nouns; past and perfect participle for verbs; inflected form, comparative and superlative for adjectives), and reference (if relevant).

Essentially new with respect to the former guides is that the selection of new entries, their orthography, and the values (the actual word forms) for the information categories per entry have an empirical basis. Evidence came from a broad coverage *50 Million Words Corpus* at INL (ca. 1600 sources, mainly 1970-1990). Both entries and word forms for an entry were included in the guide only if they met the criteria of frequency and coverage (distribution over text sources). The guide is principally corpus-based (Van Sterkenburg 1991). This implies that, for example for a particular noun, a diminutive or plural form is not listed if too few occurrences in too few sources were found in the INL corpus. .

2.4 Spelling Guide 1995

An official follow-up of the guide of 1954, *Word list of the Dutch language*, is to be published by the end of 1995. In 1994, the Dutch and Flemish Ministers of Education and Culture decided on new but not too radically changing principles for a spelling revision. Among other things, the internal consistency within the spelling system should be improved. The Dutch-Belgian

government body, *Nederlandse Taalunie (NTU)*, whose task is the promotion of the Dutch language in the broadest sense, requested INL to supply the main entries and to compile the new guide under the authority of NTU; and asked SDU to publish it in both printed and electronic form. Commissions and financial revenues were established by contracts between NTU and INL, and NTU and SDU, respectively. INL and SDU have the right to negotiate together concerning a CD-ROM publication and other spin-offs.

The guide of 1995 contains ca. 110 000 entries: the word list of 1954 from which ca. 15 000 obsolete words have been removed (ca. 50 000), the new words from the list of 1990 (ca. 30 000), and ca. 30 000 new entries. The orthography of all entries is according to the new spelling rules. The distinction between "preferred" and "allowed" spelling (cf. 2.2.) is abolished, with the exception of a few pronunciation variants only. The information categories per entry in the printed guide are essentially the same as those in the guide of 1990 (2.3.). As for the genus for nouns, the original 19 categories (combinations of male, female, and neuter) are reduced to 9 categories, accounting for the increasing tendency to consider female and male as one genus category combined with the article "de", versus neuter as an another, combined with the article "het". The database also contains the obsolete words and some additional fields per entry, among which "spelling 1990", "hyphenation 1990", "morphological category" (free vs. compound vs. derivation), "Flemish" (yes vs. non), "loan word" (yes vs. non), "year of first publication" (1954, 1990, 1995), "obsolete" (yes vs. non), and some administrative fields.

Like the guide of 1990, the guide of 1995 has an empirical basis with respect to the selection and orthography of the entries as well as the values for the information categories per entry. Empirical data were used for, for example, the choice between conflicting orthographies, such as product vs. produkt; cadeau vs. kado, scène vs. scene, mafia vs. maffia, know-how vs. knowhow, context vs. tekst. Evidence came from the broad coverage *50 Million Words Corpus*, a *5 Million Words Corpus 1994* of recent texts (17 sources, most of them dating from 1989–1994), a *27 Million Words Newspaper Corpus 1995* (1 source, 1994–1995), and other machine-readable sources available at INL[1]. Main criteria were again frequency and coverage. For some cases, the government body NTU decided on a deviating outcome, based on political considerations.

3. Spelling Guides and INL Language Resources

The INL language resources used for the spelling guides have a different status. The *50 Million Words Corpus* was compiled in the eighties when texts in machine-readable form were hardly available. Optical character recognition (OCR) was applied for converting books into electronic form. The ca. 1600 sources, for the most part dating from 1970–1990, cover a broad variety of topics. The retrieval program developed by INL allows searches at the level of word form, and for a subcorpus of 15 million words, at the levels of lemma and part of speech as well. This implies that the retrieval of frequency and coverage data for the most part concerned individual word forms rather than head words with their corresponding inflected forms in sofar as they occur in the texts. This, of course, impeded efficiency.

Since 1992, INL acquires machine-readable texts (books, magazines, newspapers, etc.) from several publishing houses on a contract basis. Due to the use of different systems for text preparation by the publishing houses, the acquired texts have different formats. Some texts are rather clean, others have a dirty proper text (i.e., full of strange characters). The encoding, if present at all, is different with respect to both the number and the types of the encoded categories. The texts were to be converted, filtered for information not relevant to this application (e.g., -usage codes), and formally harmonized to some extent, so as to make them appropriate as input for further processing and consultation. The different characteristics of the texts coming from the various publishing houses required the development of specific sets of software for handling the different text formats.

Part of these texts are included in the *5 Million Words Corpus 1994* and the *27 Million Words Newspaper Corpus 1995*. The *5 Million Words Corpus 1994* consists of several text types, most of them dating from 1989–1994, and covers a variety of topics. The *27 Million Words Newspaper Corpus 1995* covers one newspaper only, the editions dating from 1994 and 1995. These corpora have automatically been annotated for headword and part of speech, by a lemmatizer/POS-tagger developed at INL (Van der Voort et al. 1994). The retrieval program enables the user to formulate searches at the levels of word form, headword, and part of speech. Frequency and coverage data for the spelling guides could, hence, rather easily be retrieved. In order to get these data for the texts not included in the corpora (ca. 15 million words), the word forms needed still to be lemmatized and the texts to be made accessible for the purpose.

We can say that INL was equipped for the compilation of the corpus-based spelling guides by having appropriate text corpora and operational retrieval

systems at the start of the spelling guide projects. Still, considerable efforts were needed so as to make the empirical basis as large and diversified, and hence, as reliable as possible. For more detailed information on the corpus compositions and the retrieval systems, we refer to Kruyt (1995) and Van Sterkenburg & Kruyt (1996).

4. Evaluation and conclusion

Much work and time could have been saved with a higher level of harmonization in several stages of product development. In Section 3, the additional work due to the lack of standards in text preparation by the publishing houses was mentioned. Although there is a tendency towards text encoding according to the *Standard Generalized Markup Language (SGML)* standard, it will take some time before all publishers will have changed their infrastructure.

Another stage is data retrieval. In this respect, the INL textual resources have a different status (Section 3). If various text corpora are accessible on equal linguistic parameters in a uniform way, frequency and coverage data will be retrieved more efficiently. A similar argument applies to other data, of course, particularly in a multilingual environment of comparable corpora of different languages, as envisaged by the EC-funded project PAROLE.

An additional complicating factor in the spelling projects concerned the structure of the electronic spelling guide data. The electronic spelling guide of 1990 was a slightly encoded text file (2.3.). The format of the electronic file of the 1995 guide consists of a number of attribute/value pairs for each entry (e.g., for the entry "demonstratie" (demonstration) among others the pairs: flexcat (attr.): plural (value); flexform (attr): demonstraties (value)). With respect to the guide of 1990, the electronic 1995 guide has an increased number of information categories (2.4). The information of the 1990 guide was to be extracted from the text file and inserted into the new format, which involved automatic identification of specific information, making the information more explicit, and restructuring the electronic data according to the new format. For the printed edition, the publisher SDU for his part requested a selection of information categories in another, deviating format, maintaining the attribute/value pairs but with different names for the information categories and with different delimiters between the pairs. A common data structure, used by both our institute and the project partners, would have saved much programming effort.

In conclusion, we can say that future cooperation can be supported and improved by more uniform standards at the levels of text preparation, data

structure, and access to data. This does not alter the fact that the two spelling projects, as well as the earlier mentioned historical dictionary on CD-ROM (cf. 1.), demonstrate that data and facilities originally developed for internal purposes mainly may be a useful basis for product development in cooperation with commercial partners.

Note

INL offers the opportunity to consult the *5 Million Words Corpus 1994* and the *27 Million Words Newspaper Corpus 1995* by Internet, up to now for non-commercial research purposes only. In order to get free on-line access to the retrieval program developed for these corpora, a personal user agreement has to be signed. An electronic user agreement form can be obtained from our mailserver "Mailserv@Rulxho.LeidenUniv.NL". Type in the body of your e-mail message: "SEND [5MLN94]AGREEMNT.USE" or "SEND [27MLN95]AGREEMNT.USE" (without the quotes). Please make a hard copy of the agreement form, sign it, keep a copy yourself, and return a signed copy to Institute for Dutch Lexicology INL, P.O. Box 9515, 2300 RA Leiden, The Netherlands, Fax: 31 71 527 2115. After receipt of the signed user agreement, you will be informed about your user name and password. For the conditions for commercial use of INL resources, please contact the director of INL, Prof. dr. P.G.J. van Sterkenburg. If you need additional information, please send an e-mail to "Helpdesk@Rulxho.LeidenUniv.NL".

References

Herziene Woordenlijst van de Nederlandse taal. 1990. s'Gravenhage: SDU uitgeverij.

Kruyt, J.G. 1995. "Nationale tekstcorpora in internationaal perspectief". Forum der Letteren 36 (1): 47–58.

Molewijk, G.C. 1992. Spellingverandering van zin naar onzin (1200-heden). s'Gravenhage: SDU.

Sterkenburg, P.G.J. van. 1991. "Het groene boekje". Bennis, H., A. Neijt and A. van Santen (eds.). De groene spelling, 54–71. Amsterdam: Uitgeverij Bert Bakker.

Sterkenburg, P.G.J. van and J.G. Kruyt. 1996. "Dutch Electronic Corpora: their history, applications and future". Computers and the Humanities (in press).

Voort van der Kleij, J.J. van der, S. Raaijmakers, M. Panhuijsen, M. Meijering and R. van Sterkenburg. 1994. "Een automatisch geanalyseerd corpus hedendaags Nederlands in een flexibel retrievalsysteem". Noordman, L.G.M. and W.A.M. de Vroomen (eds.). Informatiewetenschap 1994. Wetenschappelijke bijdragen aan de derde STINFON-conferentie, Tilburg, 181–194.

Vries, J. W., R. Willemyns and P. Burger. 1994. Het verhaal van een taal. Negen eeuwen Nederlands. Amsterdam: Prometheus.

Vries, M. de and L.A. te Winkel. 1866. Woordenlijst voor de spelling der Nederlandsche taal. s'Gravenhage en Leiden: Martinus Nijhoff, A.W. Sijthoff.

Woordenlijst van de Nederlandse taal, samengesteld in opdracht van de Nederlandse en de Belgische regering. 1954. s'Gravenhage: Staatdrukkerij- en uitgeverijbedrijf, Martinus Nijhoff.

Woordenlijst Nederlandse taal. 1995. Den Haag: SDU Uitgevers.

European Language Resources and the Treasury of the Computerised Russian Language Fund (a Small Project Provoking a Discussion on a Big Issue)

Elena Paskaleva

Linguistic Modelling Laboratory
Bulgarian Academy of Sciences
Sofia 1113, 25a, acad.G.Bontchev str.
Tel.: +359 2713 38 41
Fax: +359 2 70 72 73
E-mail: hellen@bgcict.acad.bg.

In this presentation, I would like to draw your attention – through a small project of Central European University Research Support Scheme (CEU RSS) – to a larger issue dealing with the treasury of the linguistic resources of the Russian language.

1. Is Russian a European Language?

If a European initiative aspires to cover the treasure of European linguistic resources, it cannot afford to ignore the Russian language. Moreover, many European joint research projects include a number of Slavonic languages, and the only one that has international status is being left out. We all realise that the issue is rather of organisational than of political nature, and so we must give some serious thought to overcoming the problem. It was no accident that the earliest multi-language NLP applications turned towards the Russian language (such as the systems for machine translation in the early 60s, which started with translating "Pravda" and "Izvestija" into English).

2. The Relations between the Russian and the European CL Achievements from a Historical Perspective

During the long time we spent on both sides of the Iron Curtain, the flow of information between Russian and Western linguistic schools was seriously hindered. Obstacles existed in both directions:

a) The West was not thoroughly acquainted with the notable achievements of formal and structural models such as the model Meaning-Text. This is why, some of what is being recently introduced into wideknown models such as HPSG was developed in the above mentioned Russian model twenty years ago. The lack of knowledge of Russian linguistic achievements was more characteristic of American than of European linguistic thought. Tenier's structural syntax was the basis for Russian structural linguistics, and the developed syntactic models agreed with the achievements of the Prague school of linguistics.

b) The broken connection in the West-East direction found expression in the formalisms and in the hardware and software products used by Russian computational linguistics.

In the triad where the achievements of computational linguistics are situated, i.e., a) linguistic structures, b) formal description, and c) computational models, the Russian school was the most stable one in a), but lagged behind

in c) for reasons beyond science. The Western school had arranged its priorities in the reverse order.

Of the Eastern European countries with long traditions in computational linguistics, two are the ones which established relationships with Russian structural and formal linguistics (in its best): Czechoslovakia and Bulgaria. The former primarily because of its leading role in Slavonic language studies and the proximity of the linguistic schools. The latter because of the lexical similarity, the well established traditions in Slavonic language studies and the historically determined lack of Russophobe disposition.

The deficiency of computer resources did not prevent scientists of the Russian Language Institute from the beginning – 15 years ago – to develop a Computerised Russian Language Fund (CRLF) using sofware and hardware lagging a generation behind the Western one. A more careful look at the present CRLF archive (distributed in hyper text form, see below) convinces us that the accumulation of Russian language resources has a long way to go, which is also evident from the wide range of text processing tools used in this archive.

In the international links, however, the bilateral relations between the Russian Academies of Science and the East European countries were cancelled for financial reasons. The common projects were cancelled right at the moment when the unification of software products and the presentation of linguistic knowledge had begun.

The revival of contacts started not long ago following schemes rather different from the old ones. In the trans-European scientific cooperation, Russia was separated from the East European countries as a participant in a the special program INTAS. Presently, the funding of joint projects with participants from the West, Eastern Europe, and Russia is accidental within the frameworks of wider initiatives such as the Open Society Fund.

3. A Brief Outline of the Small Research Project

In 1995, CEU RSS (Central European University – Research Support Scheme) sponsored a small project with three participants: CRLF, Moscow; GMS (Gesellschaft fur multilinguale Systeme), Berlin, and LLM (Linguistic Modelling Laboratory, Bulgarian Academy of Sciences), Sofia. These limited resources have been granted for a project with a quite imposing title:

Application of the Data of the Computerised Russian Language Fund to the NLP Systems.

The word "apply" is a three arguments' predicate (subject- object- object). In the title above, the last two arguments need explanation. What is meant here by "data of CRLF" is the use of 10 000 vocabulary entries from Ozhegov's Dictionary of the Russian language. This dictionary is recognized by the contemporary Slavonic lexicographers as one of the best modern uni-lingual, medium-sized Russian dictionaries. "NLP systems" means MT systems like METAL and EUROTRA which execute machine translation.

The linguist researcher is attracted in this project by the requirement for unifying the parameters of the linguistic knowledge represented with the appropriate depth and scope in two real products. METAL and EUROTRA are machine translation systems dealing with syntactic information. A considerable part of the lexical entry in Ozhegov's dictionary is dedicated to the syntactic valency of the lexical item. The adjustment of the Russian data to the above systems becomes easier due also to the fact that they function as DB based on the special software UNILEX created in CRLF. These tools facilitate the transition between the two types of products with the help of an intermediate representation of, so to say, a generalized dictionary (with a temporary title ADALEX).

The Laboratory for Linguistic Modelling will participate with software products accelerating the conversion of data into the appropriate format. The design of the general ADALEX dictionary will include a permanent text support for the collection of data about the subcategorization. This requires the development of tools for textual support. The project is by no means a project for the joint work of institutions – it is simply the joint work of individual scientists towards a particular task.

The ideology of this project is in harmony with the developed trans-European projects for unification of the dictionary formats (like EAGLES and GENELEX). Thus, this project is a small model of the future real inclusion of the wealth of Russian language resources in the Europe-wide programs for linguistic technologies.

4. A Glimpse of the Actual Resources of CRLF

The modest dimensions of the resources of this project are evident when compared to the current resources of CRLF kindly presented to me by my Russian partners in the hypertext format [see(1) and (2)].

Unfortunately, the available description of the archive is in Russian, but its presentation here proves that the text- processing software products currently used in Russia are convertible in Europe. This was not the case 20 years ago.

The variety of linguistic resources in the archive is realy impressive.

Textual resources are represented in three ways: as plain text (texts in DOS and WINDOWS formats), as marked text (with manual SGML-like annotation), and as DBF files (created and supported by a special DB tool – UNILEX).

Data is stored on magnetic tapes and diskettes.

Most of the files are archived.

The collection of literary texts is striking – from Tolstoj to Brodsky.

The most precious stone in this treasury is the collection of dictionaries: most of them are in DB format. The collection includes monolingual, orthographical (general and special), syntactic, grammatical, and morpheme dictionaries of the Russian language.

The entering of the texts (in the course of 15 years) ranges from manual work to OCR processing.

In order not to seem like advertising, the above statements are accompanied by diskettes containing the description of this archive which are placed at the disposal of all participants in the workshop with the consent of CRLF.

References

L. Kolodjazhnaja. The archive of linguistic resources of the CLRF (in Russian, in electronic form).

The Russian Language Computerized Fund. Bulletin No 3, Moscow 1995 (in Russian, in electronic form).

Humor
(High-speed Unification Morphology)
A Morphological System for Corpus Analysis

Gábor Prószéky

MorphoLogic
Németvölgyi út 25.
H-1126 Budapest, Hungary
Tel.: +361 2122 429
Fax: +361 2018 355
E-mail: h6109pro@ella.hu

142

1. The Humor morphological system

Humor, a reversible, string-based, unification approach for lemmatizing and disambiguation has been used for both corpus analysis in the Research Institute for Linguistics and creating a variety of other lingware applications, like spell-checking, hyphenation etc. for the wide public. The system is language independent, that is, it allows multilingual applications; besides agglutinative languages (e.g., Hungarian, Turkish) and highly inflectional languages (e.g., Polish, Rumanian), it has been applied to languages of major economic and demographic significance (e.g., English, German, French).

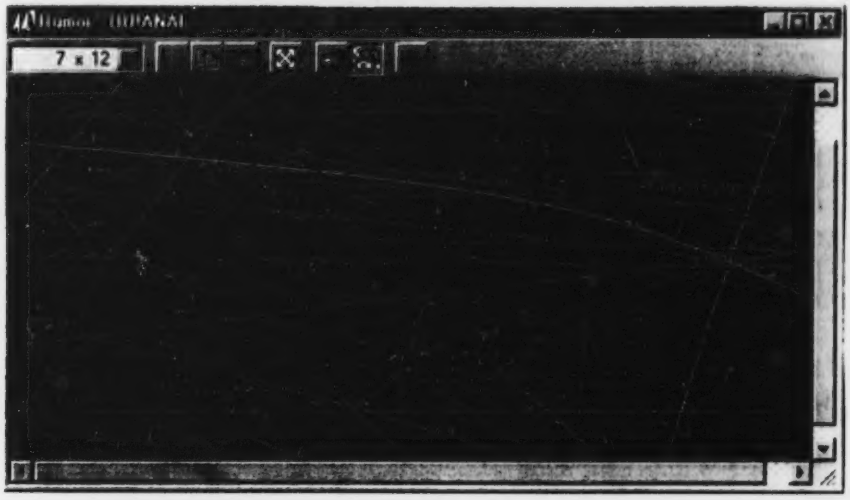


Figure 1. Hungarian morphological analysis with Humor

Humor's Hungarian version – the largest and most precise implementation – contains nearly 100 000 stems which cover all (approx. 70 000) lexemes of the *Concise Explanatory Dictionary of the Hungarian Language*. Suffix dictionaries contain all the inflectional suffixes and the productive derivational morphemes of present-day Hungarian. With the help of these dictionaries, **Humor** is able to analyze and/or generate several billions(!) of different well-formed Hungarian word-forms.

BEST COPY AVAILABLE

2. HelyesLem, the Lemmatizer

Humor has been rigorously tested both by linguists and "real" end-users of word-processing tools, namely. **Humor**-based linguistic modules have been licensed by *Microsoft, Lotus, Inso,* and other software developers. Another sort of testing of MorphoLogic's lemmatizer, **HelyesLem** (Figure 2), has also been used in every-day work since 1991 by both lexicographers and other researchers of the Research Institute of Linguistics of the Hungarian Academy of Sciences.

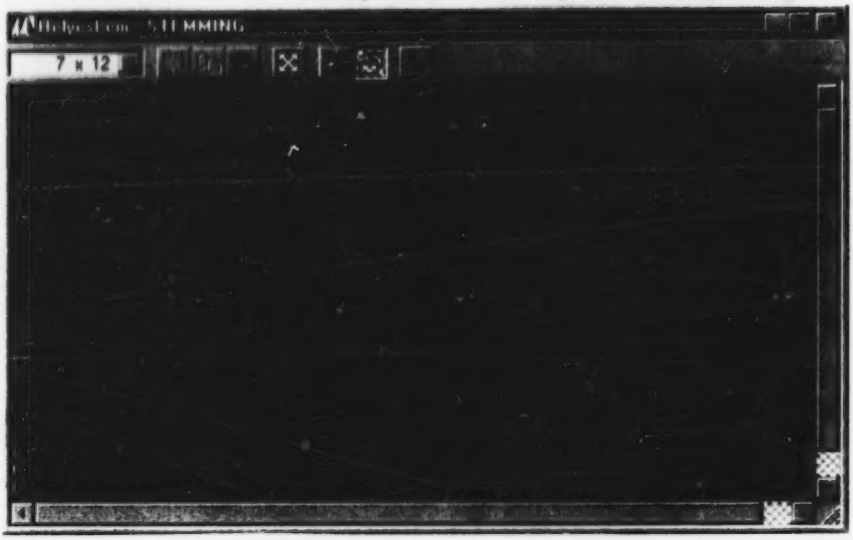


Figure 2. Stemming sample run of HelyesLem

The lemmatizer shares some of the extra features of Helyes-e?, the speller derived from **Humor**, because lexicographers need a fault-tolerant lemmatizer that is able to overcome simple orthographic errors and frequent mis-typings. Thus, it is able to treat "nonstandard" text e.g., from 19th century, when the orthographic system was not standardized as nowadays. Word-forms "auto-corrected" into the standard orthography can be properly analyzed (Figure 3).

3. Disambiguator, tagger, and parser in one tool: HumorESK

The basic strategy of **Humor** is inherently suited to parallel execution. Search in the main dictionary, secondary dictionaries, and affix dictionaries can occur simultaneously. What is more, in the near future, it is going to be

BEST COPY AVAILABLE

extended by a disambiguator based on the same strategy. This is a new parallel processing method of various levels (higher than morphology) called **HumorESK**, that is, **Humor** Enhanced with Syntactic Knowledge (Figure 4). Both **Humor** and **HumorESK** have a very simple and clear strategy based on surface-only analyses: no transformations are used; all the complexity of the systems are hidden in the graphs describing morpho-syntactic behavior.

Figure 3. Fault-tolerance of HelyesLem

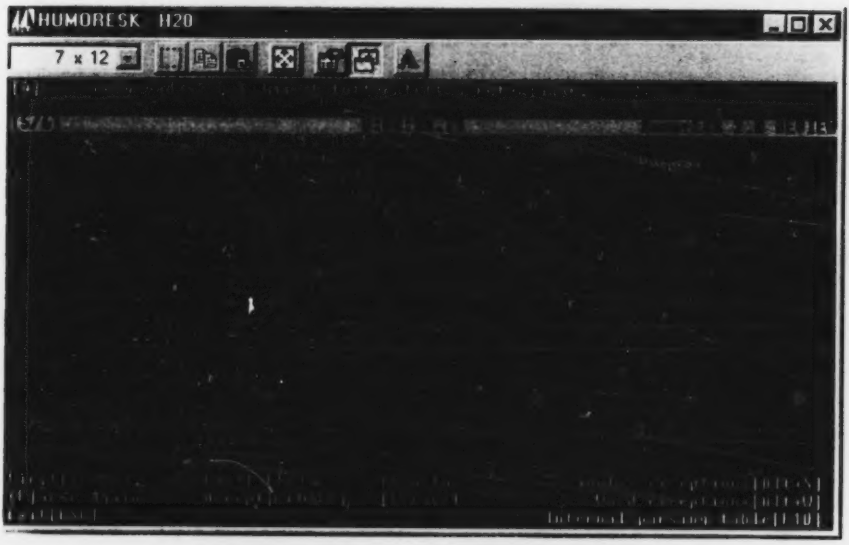


Figure 4. Parsing Hungarian with HumorESK

BEST COPY AVAILABLE

4. Languages

The **Humor** description has been made for various natural languages. Both the size of the dictionaries and the depth of the description vary, but on the basis of the experiences with them, further languages can also be described. The list of the currently existing natural languages is shown by Table 1.

East-European languages:	Ready:	Hungarian Polish Romanian
	Under development:	Bulgarian
	In preparation:	Ukrainian Czech
Non-East-European languages	Ready:	English
	Under development:	German French
	Demo versions:	Italian Latin Ancient Greek

Table 1. Languages supported by Humor

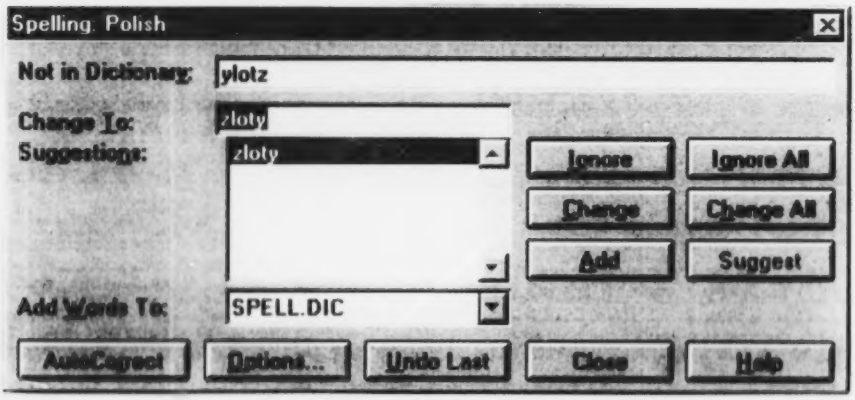


Figure 5. Polish speller based on Humor

BEST COPY AVAILABLE

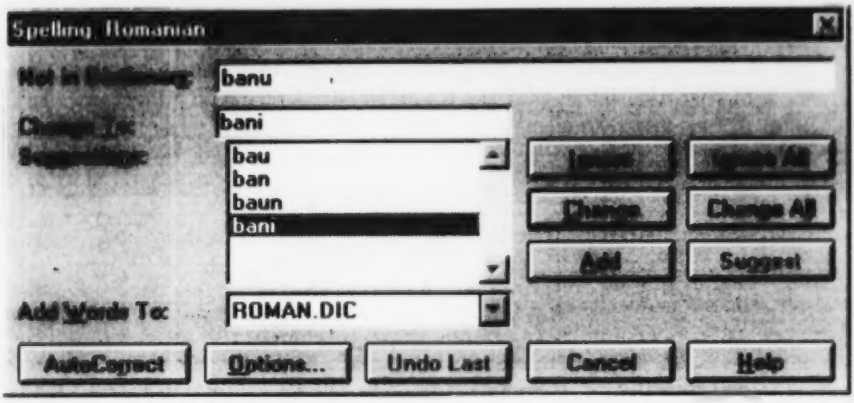


Figure 6. Romanian speller based on ***Humor***

5. Products on the Market

There are several commercially available **Humor** subsystems for different purposes: lemmatizers, hyphenators, spelling checkers/correctors and grammar checkers. They (called HelyesLem, Helyesel, Helyes-e? and Helyesebb respectively) have been built into several wordprocessing and full-text retrieval systems. Among others, Microsoft and Lotus licensed them for all of their localized Hungarian products.

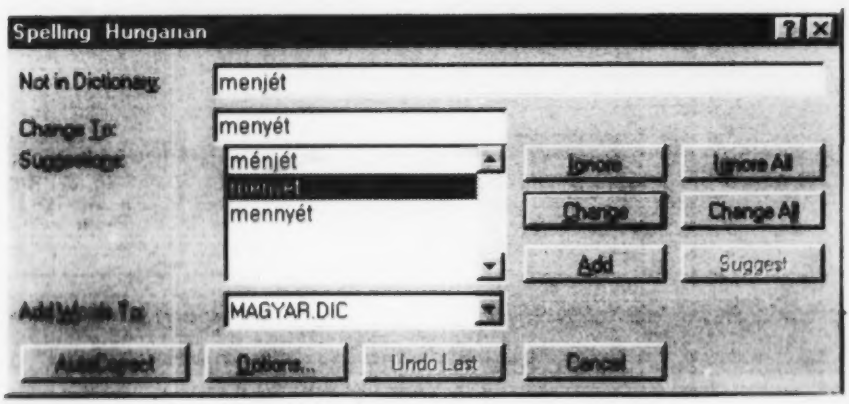


Figure 7. Hungarian speller based on Humor

BEST COPY AVAILABLE

Software products supported by **Humor**-based tools contain almost all of the Windows- and Macintosh-based word-processors (Word for Windows, AmiPro/WordPro, WordPerfect, etc.), DTP systems (Quark Xpress, Adobe PageMaker, Corel Ventura), and other applications (Excel, PowerPoint, Access, etc.) as well. Stand-alone versions work with pure text files and files in some formatting languages like RTF, PDF, SGML etc. (Figure 8)

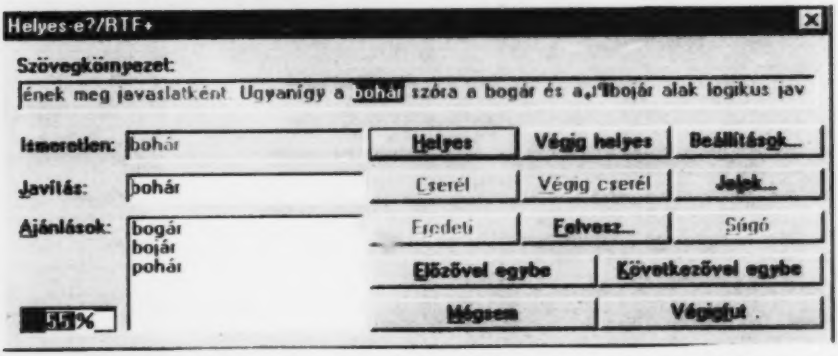


Figure 8. Application-independent (RTF) speller for Hungarian based on Humor

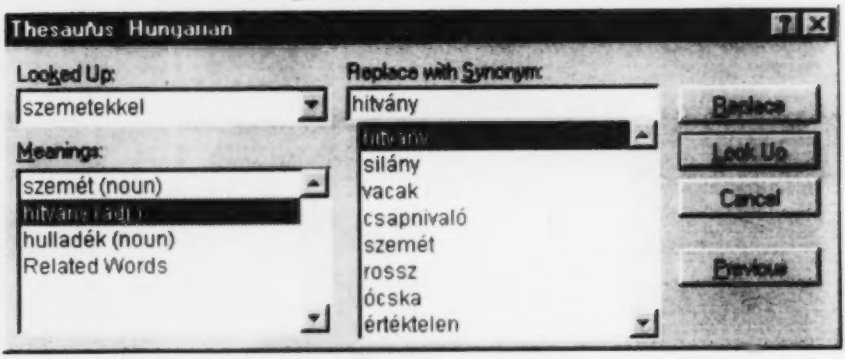


Figure 9. Hungarian thesaurus with stemming

Besides the above application there are two new tools based on the same strategy, the inflectional thesaurus called Helyette (Prószéky & Tihanyi 1992) and the series of intelligent bi-lingual dictionaries called MoBiDic. Both are dictionaries with morphological knowledge: Helyette is monolingual, while – as its name, MorphoLogic Bi-lingual Dictionary, suggests – bi-lingual. Having analyzed the input word the both systems look for the found stem in

BEST COPY AVAILABLE

the main dictionary. The inflectional thesaurus stores the information encoded in the analyzed affixes and adds to the synonym word chosen by the user. The synthesis module of **Humor** starts to work now, and provides the user with the adequate inflected form of the word in question. This procedure has a great importance in case of highly inflectional languages.

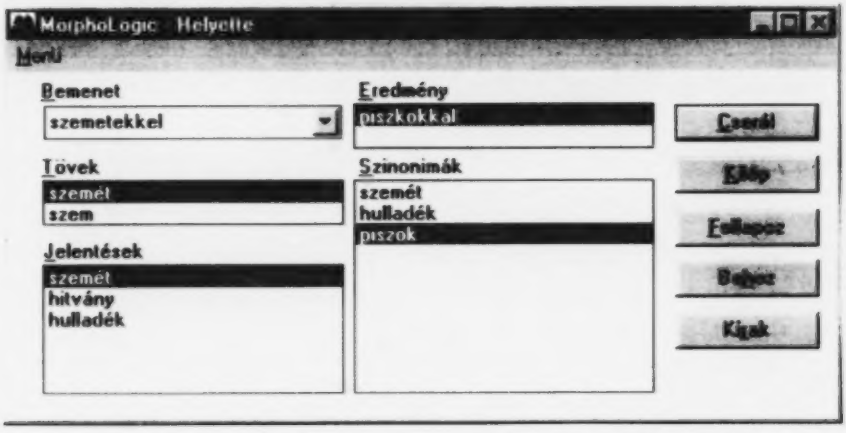


Figure 10. Hungarian thesaurus with morphological analysis and generation

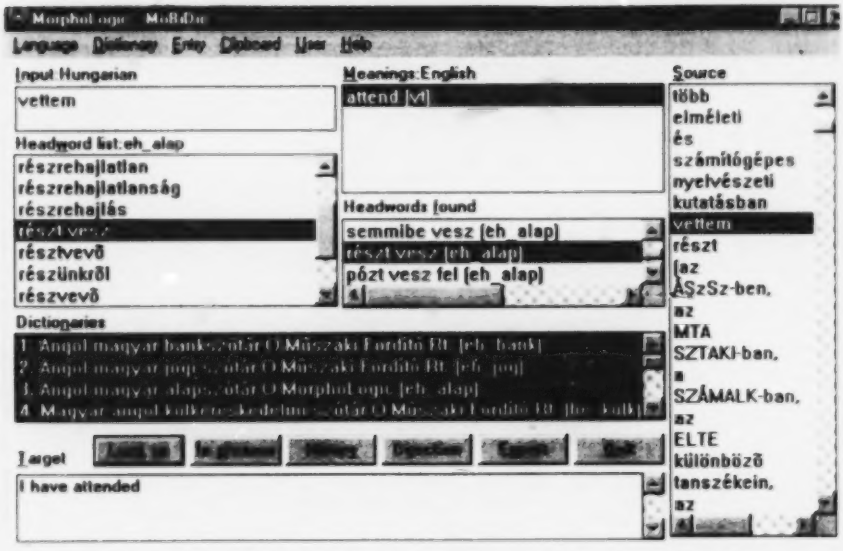


Figure 11. MoBiDic, the bi-lingual dictionary system with morphological modules

BEST COPY AVAILABLE

The synonym system of **Helyette** contains 40.000 headwords. The first version of the inflectional thesaurus **Helyette** needs 1.6 Mbytes disk space and runs under MS-Windows. The size of the **MoBiDic** packages vary depending on the applied terminological collection. E.g., the Hungarian–English Business Dictionary needs 1.8 Mbytes space.

MorphoLogic systems supporting OCR, hand-writing and speech recognition are under development. A sample run is shown on Figure 12. with the following non-segmented input:

kéremkészítsékelőjegyeiketésbérleteiketmertazállomásterületénellenőrökdolgoznak

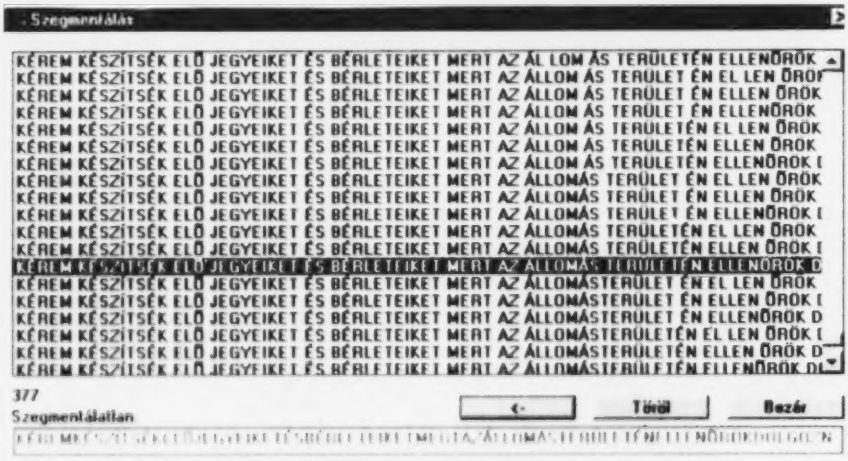


Figure 12. Segmentation of contiguous input based on ***Humor***

6. Technical details

The entire software package is written in standard C using C++ like objects. Portability to any platforms is guaranteed. The system runs on any platform where a C++ compiler can be found. MorphoLogic modules have existing implementations for DOS, Windows 3.1, Windows for Workgroups 3.11, Windows 95, Windows NT, OS/2, Macintosh, PowerMac, and Unix.

BEST COPY AVAILABLE

References

Prószéky, G. 1994. Industrial Applications of Unification Morphology. Proceedings of the 4th Conference on Applied Natural Language Processing, Stuttgart, Germany: 157–159.

Prószéky, G. and L. Tihanyi. 1992. A Fast Morphological Analyzer for Lemmatizing Corpora of Agglutinative Languages. F. Kiefer, G. Kiss & J. Pajzs (eds). Papers in Computational Lexicography – COMPLEX '92. Linguistics Institute, Budapest, Hungary: 265–278.

Prószéky, G. and L. Tihanyi, L. 1991. Helyette: An Inflectional Thesaurus for Agglutinative Languages. Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics: 473.

Prószéky, G. L. Tihanyi and M. Pál. 1994. **Humor**-based Applications. Proceedings of COLING-94, Kyoto, Japan: 1241–1244.

151

CORDON – A Joint Venture Case Study

Norbert Volz

Institut für deutsche Sprache
Abt. Sprachentwicklung in der Gegenwart
Postfach 10 16 21
D-68016 Mannheim, Germany
Tel.: +49 621 1581-437
Fax: +49 621 1581-415
E-Mail: volz@ids-mannheim.de

1. Introduction

Since I first presented CORDON at the Tihany seminar last Summer, some changes had to be made to its structure as the project became more concrete. In order to come up with a realistic, promising proposal, some early concepts had to be dismissed, others had to be revised, and where necessary, new ideas had to be followed. Still the main idea behind CORDON, to facilitate the work of the lexicographer or terminologist by using automatic procedures in order to trace and document changes in use and frequency of word forms within textual corpora, remained the same. In the following sections, I will give a short overview on the underlying concept of neologisms, the market situation and the user needs that called for the CORDON project, its scientific background, and the techniques applied for its translation into action.

2. Neologisms – Birth and re-birth of new words

Somehow, words are like people. They are born, they live, and change, and eventually they die. Language is changing constantly, responding to developments in culture, politics, society, industry, and science. New concepts emerge, requiring words for their expression.

> "What is a new word? This, of course is a question which can never be answered satisfactorily, any more than one can answer the question 'How long is a piece of string?' It is a commonplace to point out that the language is a constantly changing resource, growing in some areas and shrinking in others from day to day." (*ONW* 1991, p. v)

Sometimes, these new words are coined by the inventors of new concepts. More often, however, already existing word forms (single or multiplex) will undergo certain shifts or changes in their usage and meaning, thus enabling the integration of those new concepts in the communication process.

It is those new-born or recycled word forms that are usually called "neologisms" – in short: new, hitherto unrecorded lexical items – regardless of their meaning – that are neither proper names or acronyms, nor typographic errors, nor occasionalisms (cf. Teubert 1996).

A further distinction can be made with regard to the area in which these neologisms are emerging: On a large, common-language scale, we have the appearance of *new words* in general. In industry and science, we witness the

emergence of *new terminology* within special areas, for instance: *mouse* in the meaning "a pointing device for computers".* CORDON will deal with both areas as it is designed to cater for general as well as specific applications; therefore, we do not have to go into detail with this issue here.

The previous distinction, however, is more important: The first type is what we have already identified as "new-born" neologisms. These are mostly single words that have been newly created in a particular language, either without a previous model as is the case with most acronyms or so-called "nonce" words (e.g., *quark* in English), or by borrowing words or parts thereof such as roots, affixes, etc. from other languages such as Greek, Latin or nowadays English. To detect this type of neologism, one can countercheck the textual data against a list of already known word forms that is regularly being updated and look for suitable candidates for new word forms - a labourous and tiresome, yet fairly straightforward, procedure in itself.

The second type are the "recycled" word forms, i.e., already existing lexical items – either single words or multiple forms – whose usage and meaning have been shifted, changed, or extended. Obviously, the detection of these word forms proves even more difficult as with the first type. Just searching a full-form lexicon will not detect those lexical items that have maintained their spelling whilst changing their meaning. To detect the candidates of this type of neologism, a statistic approach, based on time-structure, frequency, and context analysis is needed.

3. The CORDON Modules

CORDON combines these two complementary approaches within a modular, step-by-step solution whose parameters can be fine-tuned to different application environments. It is centered around three main steps:
- Corpus Annotation based on full-form lexicons
- Context-based Detection Module
- Time-structure-based Detection Module

The goal of CORDON is to detect a set of "candidates" for neologisms within a textual corpus, either general language or terminology. The corpora

* However, it is clear that nowadays *mouse* in its specialised meaning is no longer to be regarded as a neologism in the strict sense, as the word became part of the everyday language of computer users. This aptly illustrates the aforementioned "life-cycle" of words.

used by CORDON will be monitor corpora according to John Sinclair's typology (Ball/Sinclair 1995), that is, collections of texts that can be acquired cheaply and without greater effort in relation to their size, e.g., machine-readable texts readily available on CD-ROM, typesetting tapes, or other electronic media. Periodical texts, for instance, newspapers or magazines are particularly useful for that purpose.* They will be linguistically annotated, i.e., lemmatised and POS-tagged, to reduce ambiguities, as this will further increase the percentage of genuine candidates for neologisms (cf. Roche 1993).

At the end of the annotation process, we now have a concordance list of unrecognised words, containing, besides proper names, typos, or special forms, the potential candidates for neologisms. To minimise "noise", i.e., remove those proper names, typographic errors, and other quasi-linguistic material and to maximise the output of genuine neologisms, a complete lexicon is to be regarded as crucial to the project.

In order to browse the texts accumulated within a month from an average daily newspaper, a lexicographer would have to examine about 40 000 word forms, of which only 2% prove to be genuine neologisms (cf. Maier-Meyer 1995, p.196). Therefore, it is necessary that proper names and alphanumeric sequences are recognised as such and not as potential candidates for neologisms. A complete lexicon will allow these sequences to be filtered, producing less noise, thus optimising the input for the subsequent statistic procedures. The lexicons used by CORDON, therefore, have incorporated lists of proper names and abbreviations. Furthermore, they include segmentation algorithms that separate alphanumeric sequences into the numeral part and the alphabetic part. The number of unknown words is considerably reduced.

For the context analysis, CORDON uses the statistics-based context analyser ENV-LINE developed at the School of English, University of Birmingham under the framework of the MECOLB project. ENV-LINE assesses the statistic significance of co-occurring words in a defined series of words and evaluates the significance of the entire series. It disambiguates words through the context and detects characteristic "example series". ENV-LINE tries to locate combinations of words that show a high degree of attraction within a file of concordance lines, tracking them down to collocates which

* At the Institut für deutsche Sprache, monitor corpora have been recently included to the corpora available within COSMAS, the IDS' corpus storage and retrieval system. There is now access to a so-called "text pool" of a regional newspaper, the *Mannheimer Morgen*. This pool consists of all newspaper articles considered for publication within the next few days and is constantly updated.

are attracted by a certain node-word. The underlying assumption being that words in a text attract each other to a certain degree, characteristic occurrences for a certain word, therefore, may be found. The program was originally conceived as a tool for lexicographers to find suitable examples of words in use; but when operating ENV-LINE, it was found that the result of the processing may also be helpful for the disambiguation of different senses of a word. The crucial point in assessing the output is the clustering of significant collocations. A word with several different meanings will show several corresponding groups of collocates; hence, a change or shift in meaning can be tracked by the software.

The core of the Time-structure-based Detection Module is a statistical analysis of the distribution of textual features within time-interval segments (Belica 1996). The corpus material used has to be suited to that purpose: Any text within the corpus has to refer to a certain sampling interval, that is, to one of a certain number of disjunct segmentation phases, according to its date.* The sampling intervals depend on the following three factors: text size, dispersivity, and homogeneity: The smaller the text size, the lesser saturated the samples will be, and the lesser significant irregularities will be detected. Dispersivity means the duration of the particular language within a selected text regardless to its position on the time scale. A newspaper article, for example, is less dispersed than a large novel. The smaller the dispersivity, the smaller the sampling intervals.

The ideal corpus for the above approach would have to be homogeneous, that is, the texts only differ in their position on the time scale, whereas all other parameters such as text type, style, genre, subject, author profile, etc., are identical. In reality, however, those ideal corpora do not exist. Therefore, the flaws caused by internal inhomogenity have to be eliminated by adjusting the corpus composition.

Basically, the statistic analysis can be divided into 10 steps:

Step 1: Definition of sampling intervals (e.g., daily, monthly, annually, or based on individual events such as elections, etc.)

Step 2: Definition of the significance factor

Step 3: Selection of the statistic features (e.g., frequency of word forms, average word length, frequency of annotates, etc.)

* The specifications for the sampling intervals have to be set according to external, problem-based criteria; for instance, the granularity, i.e. the duration and frequency of the sampling intervals, is subject to the linguistic phenomena to be analysed.

Step 4: Postulation of the zero-hypothesis: "The selected statistic feature shows a normal distribution within N intervals".
Step 5: Calculation of the statistical distribution over sampling intervals
Step 6: Validation of the zero-hypothesis for all samples within the intervals (Walsh-Test, χ^2-Test, Fisher-Test)
Step 7: Selection of the "irregular" or "abnormal" samples
Step 8: Quantification of the distribution (Kendall-Coefficient, Difference-Coefficient, Concordance-Coefficient)
Step 9: Analysis of the distribution (Context analysis, pattern recognition, clustering)
Step 10: Linguistic interpretation

At the conclusion of the above procedures, possible "candidates" for neologisms have been identified that can be further processed, for example, stored in a continuously updated database or translation memory, as is the case with the CORDON Demonstrator application.

The front-end used for CORDON will be based on the user interface designed for the machine-based translation system STA that has been developed by one of the consortium partners, THAMUS Spa. As this software product is based on Microsoft Windows®, large portability and affordability of the final CORDON product is guaranteed.

4. The CORDON Consortium

The CORDON consortium consists of five academic and five industrial partners or subcontractors. The academic partners are:

BIR University of Birmingham, School of English, Birmingham, U.K.
CIS Centrum für Informations- und Sprachverarbeitung der Universität München, Munich, Germany (Subcontractor)
GOT University of Gothenburg, Department of Swedish, Gothenburg, Sweden
IDS Institut für deutsche Sprache, Mannheim, Germany
PAR Laboratoire de Linguistique Informatique Université Paris XIII, Paris, France

The industrial partners are:

CBD COBUILD Ltd., Birmingham, U.K.
GCP GECAP Gesellschaft für technische Information mbH, Munich, Germany

KFT Krupp Fördertechnik GmbH, Duisburg, Germany (Subcontractor)
MDD La Maison du Dictionnaire, Paris, France
THA THAMUS, Consorzio per la Linguistica Computazionale, Salerno, Italy (Coordinator)

The project duration will be two years with an assigned manpower of 184 person-months. Estimated project costs are 2061 kECU of which 1486 kECU are expected to be EC-funded.

THA will be responsible for the coordination of the project. As a company with proven expertise in the area of language engineering, they will ensure a smooth and efficient overall organisation of the project under an experienced and professional management. They will organise the semiannual meetings and will also be responsible for information flow and decision enforcement among the partners and accounting of the project.

KFT with their large experience in LR applications on terminology and translation will be responsible for the specification of validation criteria in order to ensure maximum response to user needs and to minimise the marketing risks for the final product. For the tools and resources available within the CORDON consortium, the main evaluation criteria are their generic applicability as modules for the tasks encompassed by the project, their conformance to already existing standards, and their response to user needs as denoted by the industrial user community. For the CORDON demonstrator itself, the validation process will help to calibrate parameters, specifications and algorithms according to the characteristic features of a limited number of applications, i.e., lexicon development, translation, and terminology management.

Assessment and evaluation are undertaken at defined stages ("evaluation breakpoints") within the overall project (months 12 and 18), that is, after completion of the corpus and functional specifications. In addition to these evaluation stages, a "dummy version" of the CORDON demonstrator is made available already at a very early stage of the project to allow for effective verification and quick response to user needs, especially in the areas of ergonomy and front-end performance. This "dummy version" will successively be upgraded or replaced by the inclusion of more modules.

From the very beginning of the project, one of the industrial partners, CBD, will be responsible for the establishment of an Industrial Interest Group (I.I.G.) which is to be involved in every stage of the project in order to ensure verification and alignment to the needs of the user community. The I.I.G. will represent the typical prospective users of CORDON. It will include end users such as translators, producers of technical or scientific texts, etc. as well as processing users such as terminology providers, and lexicon and

dictionary publishers. The I.I.G. will meet on a semiannual basis, allowing on the one hand the users to influence and verify results and performance during the project work while on the other hand enabling future users to make themselves familiar with CORDON while the project is still running. This has two main advantages: It will speed up the start of usage of the final software product and will also reduce the associated marketing risks once the project is finished. The final CORDON demonstrator, therefore, will be the result of a continuous process, involving potential users at every stage.

5. The Market Situation

During the last decades, the progress within the area of information technology in general and language engineering in particular has shown a considerable impact on the acquisition, storage and maintenance of LR, and especially of corpus-derived lexical resources. Vast amounts of linguistic data material can now be accumulated at a fraction of the price of several years ago; and today, even small desktop PCs allow the access to CD-ROM databases many megabytes in size. As a result, today many more organisations, industrial enterprises, academic institutes, and private persons than ever before work with machine-readable textual data.

Machine-readable corpora can be described as "raw resources". They are used for the production of lexicons, which in turn form the central components for various other language applications such as translation memories, terminology databases, machine-readable dictionaries, etc.

Those raw resources are also capital investments. Unfortunately, textual resources become outdated rather quickly with natural language being a highly unstable communication system subject to changes affecting society, science, and technology. Therefore, excellent material is of utmost importance, as competition on the LR market is hard, and especially small and medium-sized enterprises (SMEs) that build the majority of Europe's LR and LT industry must ensure a product of high degree quality. Maintaining and enforcing the actuality of the resources used is vital for both the supplier and the end user of lingware products: It is important for the supplier as it will maintain the market value of the product. In other words: Old, out-dated applications and resources will not sell. It is important also for the user: Only actual, up-to-date material will guarantee a high degree of quality and competitiveness of the end product or service.

6. Conclusion

In this project, the European Community is expected to cover the major part of the costs. This is often the case with projects that are concerned with the development of innovative technologies. However, the participation and financial support from industrial partners is also very important and will become increasingly important in the future. Still, commercial enterprises will only invest in areas where they can expect to gain profit.

Therefore, a close cooperation between industrial and the academic partners is vital. The aim has to be the development of competitive products for the European and global LR and LT market. The academia will provide the resources and expertise needed, including generic software. The industrial partners will ensure that the final product meets the demands of the user community. CORDON shows that the above cooperation is possible and can be used to provide sophisticated generic tools that will facilitate communication among Europe's various languages.

So far, Europe still has a leading edge in LR and LT applications. Only cooperation between public domain research and industry will maintain and strengthen this leading position on the international market.

7. References

Ball, J. and J.M. Sinclair. 1995. "Corpus Typology". EAGLES working paper. Electronic document on the University of Birmingham FTP server. University of Birmingham.

Belica, C. 1996. "Analysis of Temporal Changes in Corpora". In: International Journal of Corpus Linguistics, Vol. 1,1 (forthcoming).

Guenthner, F. and P. Maier. 1994. Das CISLEX-Wörterbuchsystem. München: Centrum für Informations- und Sprachverarbeitung der Universität München (CIS-Bericht 94–76).

Maier-Meyer P. 1995. Lexikon und automatische Lemmatisierung. München: Centrum für Informations- und Sprachverarbeitung der Universität München, (CIS-Bericht 95–84).

Tulloch, S. (ed.). 1995. The Oxford Dictionary of New Words. Compiled by Sara Tulloch. Reprinted with corrections. Oxford, New York: Oxford University Press.

Roche, E. 1993. "Analyse syntaxique transformationelle du français par transducteurs et lexique-grammaire". These de doctorat en informatique, Université Paris 7, Paris.

Teubert, W. 1995. "Neologie und Korpus". In: Korpus und Neologie. (To be published)

EVA - A Textual Data Processing Tool

(Software Demonstration)

Primož Jakopin

Institute for the Slovenian Language ZRC SAZU
Novi Trg 4
SLO-61000 Ljubljana
Tel.: +386 61 1256 068
Fax: +386 61 1255 226-253
E-mail: primoz.jakopin@uni-lj.si

At the end of 1982, EVA developed out of an experiment, to test if a humble personal computer of the day and place, Sinclair ZX Spectrum, could be put to reasonable use for text and data processing instead of a mainframe computer. The positive answer lead to a full-blown text editor with integrated data base and graphics facilities in 1985, which in 1986 has been ported to the ATARI ST computer family (named STEVE). The DOS version has been in use since 1991; the Windows NT/Windows 95 version is under development. EVA has served as a software tool for processing of a sizeable amount of textual material and for preparation of various dictionaries in the Slovenian academic environment, where UNIX-based work stations have found very little response in the humanistic domain so far.

From the start, the program has been designed to be as flexible as possible, in order to allow the user to accomodate his own needs and situations with little external support. It is more or less self contained, with its own keyboard table, screen characters, DTP mode, graphics editor and OCR facility. To answer the needs for wide character sets such as UNICODE, a capability to process 8-bit and 16-bit characters in the same file has been introduced in 1993. If a line of text contains only characters with codes below 256, it is stored as a string of 8-bit characters in RAM as well as on disk. If, on the other hand, it contains one or more characters with codes above 255, it is stored as a 16-bit entity. In figure 1 the list of the first 1209 characters from the EVA character set is given.

Of special interest are codes 13 to 28, which modify the look of printed characters and sets 279 to 310, 327 to 354 and 365 to 369. The latter three groups represent the upper-, lower-, and through- diacritical characters which can all be combined with any other character to create characters, which are not part of the EVA character set.

A large set of data base functions includes general purpose routines such as sequential or indexed sorting and searching, as well as more specialised functions such as splitting of text into sentences, word or fieldwise translation and markup, or the computation of entropy. The majority of interfaces for import and export to other software is dual: ASCII file format (with user-definable filtering), RTF format (again with an open filter, for interface with Microsoft Word), WordStar file format, STEVE file type, open data record interface and the PCX picture format. For input there are also WordPerfect and TIFF interfaces.

Currently EVA is also used in production of a lemmatization dictionary of Slovenian, based on the Dictionary of the Slovenian Literary Language (93 151 headwords). So far, entries for nouns (54 522 lemmas to make 468 281 word forms) and adjectives (22 961 lemmas and 277 831 word forms) have been completed.

Figure 1: EVA character set – characters with codes from 0 to 1208

BEST COPY AVAILABLE

On-line Access to Linguistically Annotated Text Corpora of Dutch via Internet

J.G. Kruyt, S.A. Raaijmakers, P.H.J. van der Kamp, R.J. van Strien*

Institute for Dutch Lexicology INL
P.O. Box 9515
NL-2300 RA Leiden
Tel.: +31 71 527 2270
Fax: +31 71 527 2115
E-mail: kruyt@ruLxho.Leidenuniv.nl

* The first two authors are working at the linguistic Language Database Department, the other two at the Electronic Data Processing Department.

1. Corpus development at the Institute for Dutch Lexicology INL

The Institute for Dutch Lexicology INL is a research institute subsidized by the Dutch and Belgian governments. Corpus development at INL dates from the mid-seventies. Up to 1990, the INL text corpora were developed for lexicographical purposes mainly. Presently, they are used for a broad range of research and applications (cf. Van Sterkenburg and Kruyt in press). A recent example is the official Dutch spelling guide published in 1995, which is based on INL text corpora (Kruyt and Van Sterkenburg this volume).

INL text corpora of present-day Dutch include two linguistically annotated corpora which can be consulted via the international computer network Internet: the *5 Million Words Corpus 1994*, which covers a variety of topics and text types, and the *27 Million Words Newspaper Corpus 1995*. A corpus of ca. 30 million words, with varied composition and with extended linguistic encoding, will be ready for similar use in spring 1996. The present paper reports on the former two corpora already accessible via Internet.

2. Characteristics of the corpora

The *5 Million Words Corpus 1994* contains seventeen text sources, most of them dating from 1989-1994. The texts are classified along the parameters publication medium (book, newspaper, magazine, written-to-be-spoken) and topic (politics, journalism, leisure, linguistics, environment, business and employment). The *27 Million Words Newspaper Corpus 1995* covers one newspaper only, with editions dating from 1994 and 1995.

The texts of both corpora were acquired in machine-readable form, on a contract basis with the provider. The contract specifies the conditions of use, taking into account issues of copyright. Permission has been obtained for use of the texts in this particular application. After some preprocessing (Kruyt and Van Sterkenburg this volume), the texts were input for automatic linguistic encoding. Part of speech (POS) and headword were automatically assigned to the word forms in the electronic texts by lemmatizer/POS-taggers developed by INL. The lemmatizer/POS-tagger DutchTale (Van der Voort van der Kleij et al. 1994) was applied to the *5 Million Words Corpus 1994*. An improved version of this program has been used for encoding the *27 Million Words Corpus 1995*. This new version, *DutchTale II*, uses separate rule files, which allows for easy inspection and modification of the implemented linguistic knowledge. The addition of a more elaborate morphological module, incorporating, amongst others, compound analysis, has resulted in an increased number of analysable

tokens (individual word forms). Supplementary disambiguation rules have contributed to a higher precision of disambiguation. *DutchTale II* is implemented in C and runs on the institutional VAX.

Most of the data has not been corrected, neither at the level of the proper text, nor at the level of POS and headword.

3. Retrieval facilities

The linguistically encoded texts were loaded into an on-line retrieval system developed by INL. Queries may address the whole corpus, or a subcorpus defined by the user. Parameters for the definition of subcorpora are text source, topic and publication medium for the *5 Million Words Corpus 1994*, and year and month of publication for the *27 Million Words Newspaper Corpus 1995*. The system allows the user to search for single words or word patterns, including some, rather primitive, predefined syntactic patterns which can be customized by the user. Search definitions may include references to word forms, POS and headwords, both separately and in combination by use of Boolean operators and proximity searches. Some examples of queries are:

(Boolean) lemma='hongar*' and not pos='a'

This query searches for lemmas compliant with the pattern 'hongar*' (the asterisk serves as a wildcard) with part of speech not equal to 'a' (adjective).

(proximity search) lemma='president+koning*+staatshoofd'][?¦0..3]<'PP'>

In this query the '+' acts as the Boolean operator OR. So, the query searches for lemmas compliant with either 'president' OR 'koning*' OR 'staatshoofd' followed by a 'PP' (prepositional phrase) within at most 3 arbitrary words.

The present user interface appears to be rather complex, in particular for unexperienced users, due to the high degree of formalism. During the seminar, a more elegant user interface was demonstrated, containing a reduced-formalism interpreter. The interpreter allows the user to enter his query with a less elaborate notation. The retrieval engine, however, works with the complex formalism, so translation is necessary. With this interface, the latter example can be entered as:

le=president or koning* or staatshoofd dist 3 ? cat=PP

Also, a prototypical natural language interpreter is under development. This interpreter accepts queries in plain Dutch. An example is:

geef mij alle lemmas die niet op "heid" uitgaan
'give me all lemmas that do not end with "heid"'
which is translated into: not lemma='*heid'

The natural language interpreter will operate in tandem with the reduced-formalism interpreter; if the natural language interpreter fails to comprehend the query, the user will have to address the reduced-formalism interpreter. This interface will be implemented in the *30 Million Words Corpus* planned for 1996.

Output data of the retrieval system include intermediate tables with the possibility of selecting specific items (word forms, lemmas and POS with their frequencies), and ultimately a series of concordances of the searched item(s) (i.e. the searched term(s) in the local context), with a user-defined context size. Concordances can be sorted by the user along several parameters. A few concordances for the Boolean and proximity searches formulated above are:

27 Million Words Newspaper Corpus 1995
For the INL, Leiden 11 16 1995, version 1.01

NRC_NOV_94* terracotta vazen uit	**Hongarije.**	„Ik heb een grote boerderij, di
NRC_NOV_94* s collega's uit Polen,	**Hongarije,**	Tsjechië, Slowakije, Roemenië en
NRC_NOV_94* se diplomaat. Polen en	**Hongarije,**	die al formeel het lidmaatschap
NRC_NOV_94* ehouden met Estland en	**Hongarije,**	maar daar heb ik nooit het predi
NRC_NOV_94* munistische landen als	**Hongarije**	en Bulgarije. De grondwet definie
NRC_NOV_94* ë, Slowakije, Polen en	**Hongarije.**	Terugkijkend waren er al in sept

...

<PREV>/<NEXT>=previous/next page, <8/HELP>=help

27 Million Words Newspaper Corpus 1995
For the INL, Leiden 11 16 1995, version 1.01

NRC_NOV_94* es van commissaris der	**koningin in Zuid-Holland** en secretaris-gene
NRC_NOV_94* e. Daarvoor zou men de	**president van de Europese Beweging** in Frank
NRC_NOV_94* ordeel Jakarta, 1 Nov.	**President Soeharto van Indonesië** acht de ad
NRC_NOV_94* aangespannen tegen het	**staatshoofd wegens onbehoorlijk bestuur.** He
NRC_NOV_94* Hafr Al-Baten, 1 Nov.	**Koning Fahd van Saoedi-Arabië** heeft toegege
NRC_NOV_94* i Boldyrev, dat hij de	**president herhaaldelijk 'heeft ingelicht** ove
NRC_NOV_94* lag. Daarop verdedigde	**president Jeltsin Gratsjov in zeer lovende**
NRC_NOV_94* en woordvoerder van de	**president in Kaapstad** ondubbelzinnig had on

...

<PREV>/<NEXT>=previous/next page, <8/HELP>=help

Due to copyright restrictions, a limited number of concordances can be transferred to the user's computer by e-mail. It is not allowed to transfer complete texts or substantial text fragments.

The retrieval system is running on a VAXstation 4000/90A, under OpenVMS 6.0. It was developed in VAX Pascal, using the VAX SMG-routines for screen handling (cf. Van der Voort van der Kleij et al. 1994). The more elegant user interface was developed with TPU, a user-extendible text processor for VAX systems. The INL VAXstation is a multi-user computer concurrently used by many colleagues working on various INL projects. In order to restrain the guest users from accessing data to which they are not authorized, they are locked up in so-called captive accounts, a feature of the VAX/OpenVMS operating system. These accounts allow them only to run the retrieval program(s) for which they have signed the corresponding user agreement(s) (see below). Furthermore, all actions of the guest users are stored in logfiles, which are used for internal statistics and security reports. Additionally, an analysis of the list of queries will be used for enhancing the retrieval system.

4. Access to the corpora.

Consulting the corpus is free of charge for non-commercial, research purposes, provided that a personal user agreement is signed. The user agreement includes the conditions of use. For academic teaching purposes, special arrangements are possible after consultation with the first author or the director of INL, Prof. dr. P.G.J. van Sterkenburg. The conditions for commercial applications are to be discussed with the director of INL.

To gain access to the corpus, an electronic user agreement form is to be obtained from our mailserver Mailserv@Rulxho.Leidenuniv.NL. Type in the body of your e-mail message: SEND [5MLN94]AGREEMNT.USE or SEND [27MLN95]AGREEMNT.USE, for the *5 Million Words Corpus 1994* and the *27 Million Words Newspaper Corpus 1995*, respectively. Please make a hard copy of the agreement form, sign it, keep a copy yourself, and return a signed copy to: Institute for Dutch Lexicology INL, P.O. Box 9515, 2300 RA Leiden. After receipt of the signed user agreement, you will be informed about your username and password. Note that the use of a VT 220 (or higher) terminal, or an appropriate terminal-emulator (e.g. Kermit) is recommended. If you need additional information, please send an e-mail message to Helpdesk@Rulxho.Leidenuniv.NL.

References

Kruyt, J.G. and P.G.J. van Sterkenburg. "A New Dutch Spelling Guide". This volume.

Sterkenburg, P.G.J. van and J.G. Kruyt. In press. "Dutch Electronic Corpora: their history, applications and future". *Computers and the Humanities.*

Voort van der Kleij, J.J. van der, S. Raaijmakers, M. Panhuijsen, M. Meijering and R. van Sterkenburg. 1994. "Een automatisch geanalyseerd corpus hedendaags Nederlands in een flexibel retrievalsysteem". Noordman, L.G.M. and W.A.M. de Vroomen (eds.). *Informatiewetenschap 1994. Wetenschappelijke bijdragen aan de derde STINFON-conferentie,* Tilburg, 181-194.

Tagging a Highly Inflected Language (Non-English, Non-Latin Alphabet, No Morpho Component)

Elena Paskaleva, Bojanka Zaharieva

Linguistic Modelling Laboratory
Bulgarian Academy of Sciences
1113 Sofia, 25A Acad. G. Bontchev str.
Tel.: +359 2 713 38 41
Fax: +359 2 70 72 73
E-mail: hellen@bgcict.acad.bg

1. Use of the Product

1.1 General Limitations

The SUPERLINGUA system here presented carries outpresented here implements a special kind of tagging which is far behindlow the power of modern tagging systems. However, as it is pointless to argue that only heavy weaponry is used in a war, in the NLP battle there are also situations where applying all the technology for the processing of linguistic knowledge in corpora is impossible. These situations are of diverse nature, but some of their combinations include:

a) a new, virgin language is processed for which NLP technology has not been developed, or

b) the available technology is inaccessible due to the incompatibility of platforms (hardware or software), and also of institutions (the latter being the hardest to overcome). This partially accounts for the paradox contained in the title: highly inflected language with no morphological component.

1.2 Some Local Limitations

In marketing research, the orientation towards a particular target user is an essential prerequisite for the creators of general software. For NLP products and their application in a highly specialised area, such research is even more important. In countries with a young market economy and serious economic problems, such research is literally a must.

A specific restriction that has to be made is typically Bulgarian and deals with the hardware resources available to the potential users of the system in question – linguists and translators. The marketing research available on computer technology in Bulgaria does not refer to that user-group directly. In a rough outline of the distribution of hardware, banks – currently the most prosperous users – and the Bulgarian linguists, translators, and philological educational institutions obviously occupy the two extremes of the ladder of wealth.

There are many users in Bulgaria who cannot take their pick among different WINDOWS versions (3.1, 3.11 or WINDOWS 95 is unavailable because of their configurations of AT-286 with memory of 2 MB or less) because they do not have powerful enough computers.

In SUPERLINGUA, the chosen hardware configuration and the software platform are crucial for the purpose of outgrowing the boundaries of Research and Development products and discussing, if not an industry, then at least a service to the mass user. This is especially valid for the countries of Eastern Europe with their particular level of computerisation before as well as after the Great Change. The system presented here also has its place in the system of criteria for evaluation of the applicability of a computer product.

1.2.1 Software configuration of SUPERLINGUA and characteristics of the program design

The system claims to really be a working one and to be targeted at the real, i.e., the poor Bulgarian user in question; thus, it has been developed in the DOS environment. With the facilities of that environment, all qualities of the user interface have been developed which are natural functions of every WINDOWS application, including the least extravagant ones.

1.2.2 User and data-type orientation

The system has been designed with the linguist born in mind as the specific user whose hardware deficiency is accompanied by a deficiency of knowledge and skills in computer processing practice. The system is oriented towards the peculiarities of the processed material – large text files and low speed of manual tagging. Options have been developed for that purpose, allowing interruption and resumption of processing with retaining of the intermediate results as well as division of large files into portions. In the process of this arduous work, fool-proof facilities and warning messages have been designed to assist the exhausted or/and artless linguist.

1.2.3 Language orientation

The degree of universality of a certain procedure for extracting linguistic knowledge is determined by the depth of the representation of that knowledge. The only level where procedures can be universally applicable is the text-string level (the normal level of text editors). Even here, the existing discrepancies in character-set representation bring in language-specific

elements. In the DOS environment and the employed DB language (CLIPPER), versions have been developed for Bulgarian, English, and Latin – languages which distribute ASCII space uninterruptedly. A separate version is being developed for Russian texts situated in the Russian coding area, i.e., processed and used in Russia.

1.2.4 Current developments in the software configuration

For the wealthier users, the development of a WINDOWS version of SUPERLINGUA is being prepared with the resources of ClipWin in which many of the specific user-oriented and quasi-WINDOWS features of the system will be included in the very design of the program environment.

2. Functions of the System

The basic function of SUPERLINGUA is the *grammatical tagging* of a text: it carries out the steps in a grammatical analysis of the words in it. These steps as well as all supplementary and accompanying procedures are carried out by the basic modules of the system as follows:
1. Preprocessing module; **2.** Lemmatisation; **3.** Identifying the characteristics of the lemma; **4.** Identifying the characteristics of the word forms of a lemma.

2.1 Basic and Secondary Functions

The basic items and operations of the above mentioned modules allow the application of supplementary procedures: the results, although foreign to tagging itself, have their own significance not only as linguistic operations but as a basis for other important NLP applications. Examples of such an application are:

a) The results of pre-processing of a sufficiently large linguistic database used in an automatic sentence-segmentation and sentence-alignment;
b) The sum of linguistic data obtained from lemmatisation (the transformation of the text to its vocabulary) gives the core of various statistical research and all of the formats of text vocabularies.
c) The developed functions of allowing constant text-support (with editing and revising facilities) also allow for the establishment of all possible types of concordances.

2.2 Principles of Structuring of the System's Linguistic Design

The noble intentions of the SUPERLINGUA's creators are: to provide the linguist with all he wants and is able to receive on the language level within system's reach, in team-work with it. These ideas are materialized in the main principles of the linguistic design, namely:

a) maximum extraction of linguistic information from the current level of presentation and processing;
b) compensation for the lack of knowledge within the system by means of a user-friendly system of menus extracting the user's linguistic knowledge.
c) possibilities for error correction at any stage of the processing;
d) the use of contextual support at all stages and levels of the representation.

The system's linguistic design modeulates the grammatical tagging of the text. This operation which means juxtaposition of the total of text units to the total of those units along with the grammatical information attached to them. There are various degrees in the automatization of tagging, and the two extremes are marked by:

a) completely manual tagging where, for every running word in a sentence, the result of grammatical analysis is entered; therefore, the computer is simply a text-editing tool in that process;
b) a fully-automated tagging which presupposes the availability of a morphological analyser and a disambiguator (which are rule-based or statistical).

Neither can the former operation be completely manual, nor can the latter be fully-automated. The degree of automatization of SUPERLINGUA is the maximum possible automatization of manual tagging where morphological analysis and disambiguation are manual operations assisted and facilitated by the system. Assistance is offered primarily in the structuring of grammatical knowledge.

3. Units and Operations in SUPERLINGUA

In the transition from the original (untouched, virgin) sequence from the input text file (T) to the tagged text (TagT), the following portions of linguistic knowledge are involved, and are considered below according to the levels of knowledge and according to their nature: actual linguistic unit, and information about them.

3.1 Knowledge of the Text Units

In SUPERLINGUA's knowledge of the textual level, the basic unit of representation and processing is the running word (**W**). **W** is a symbol sequence consisting of at least one letter. Other symbol constituents of **W** can be:

- letters: a) belonging to the processed language or to the foreign one;
  b) small or capital;
- figures;
- punctuation marks;
- graphic symbols.

Some criteria based mainly on the formal structure of the concrete **W** distrubute the set of all **W**s into five subsets:

1. regular word forms (**WF**),
2. abbreviations (**Abr**) or
3. capitalized abbreviations (**CapAbr**),
4. proper names (**PropN**), and
5. textual garbage (**Garb**).

The distribution of **W** into the five subsets above is fulfilled by the *preprocessing* module of SUPERLINGUA (Fig. 1). In the preprocessing, the distribution of all **W**s into the sets **WF, PropN, Abr, CapAbr,** and **Garb** is executed in two ways – fully automatically and in a dialogue with a contextual help for the ambiguous cases **W/Abr** and **W/PropN** located on the sentence boundary.

The elements of the first subsetunit (**WF**) provides the connection with the next level of linguistic knowledge – the *morphological* one. This first order distribution is the bridge thrown from the plain reality of to the linguistic representation of the *text* sequence to the ordered reality of *vocabulary* units. In the construction of **WF** vocabulary, the ordering is not a proper linguistic one – it is reduced to a simple alphabetizing and exclusion of repeated elements. The only linguistic knowledge used in a system's dialog is the user's knowledge (supported by the contextual help) in the disambiguation **W/Abr** and **W/PropN** for the boundary cases.

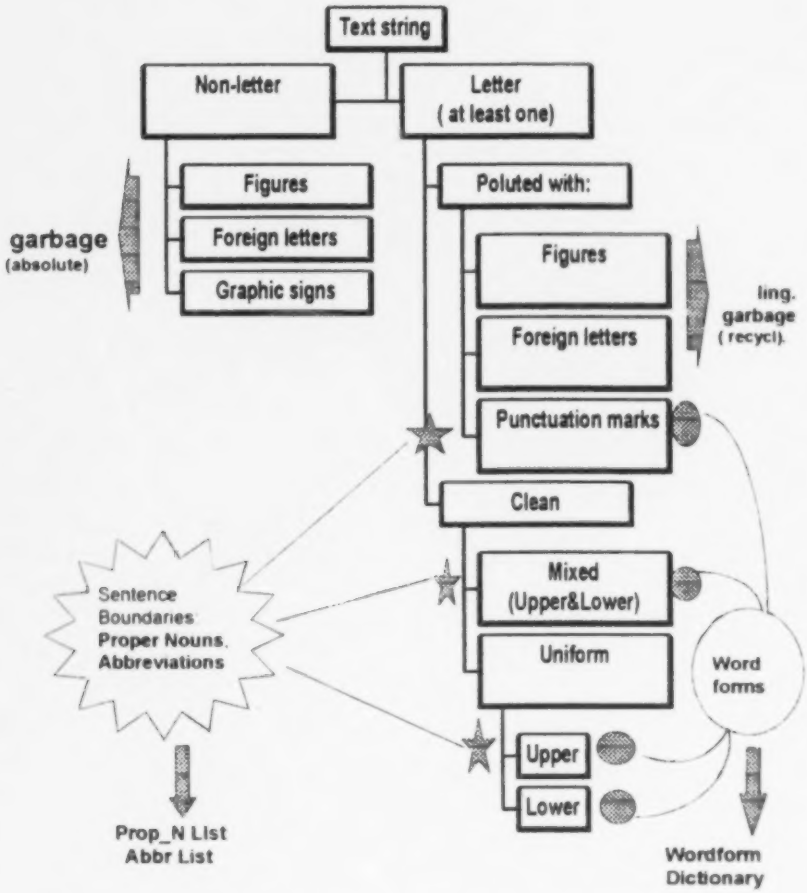


Figure 1. A first order distribution of the text sequence in the preprocessing module

3.2 Knowledge of the Linguistic System

In the system's knowledge of the morphological level of representing the linguistic units, a running word can be a word form (**WF**) or a lexeme (**L**). In the paradigmatic presentation, **WF** and **L** are *variant* and *invariant, particular* and *general*. Two corresponding types of grammatical knowledge are formed

BEST COPY AVAILABLE

during the extraction of linguistic information from the text: the former is *lexical* (**LexInfo**), and the latter is *grammatical* proper (**GramInfo**). **LexInfo** is included in general dictionaries in each *lexical* entry, and the **GramInfo** represents the result of the *grammatical* analysis of each word form obtained through either the Morpho-component or manually (see Figure 2).

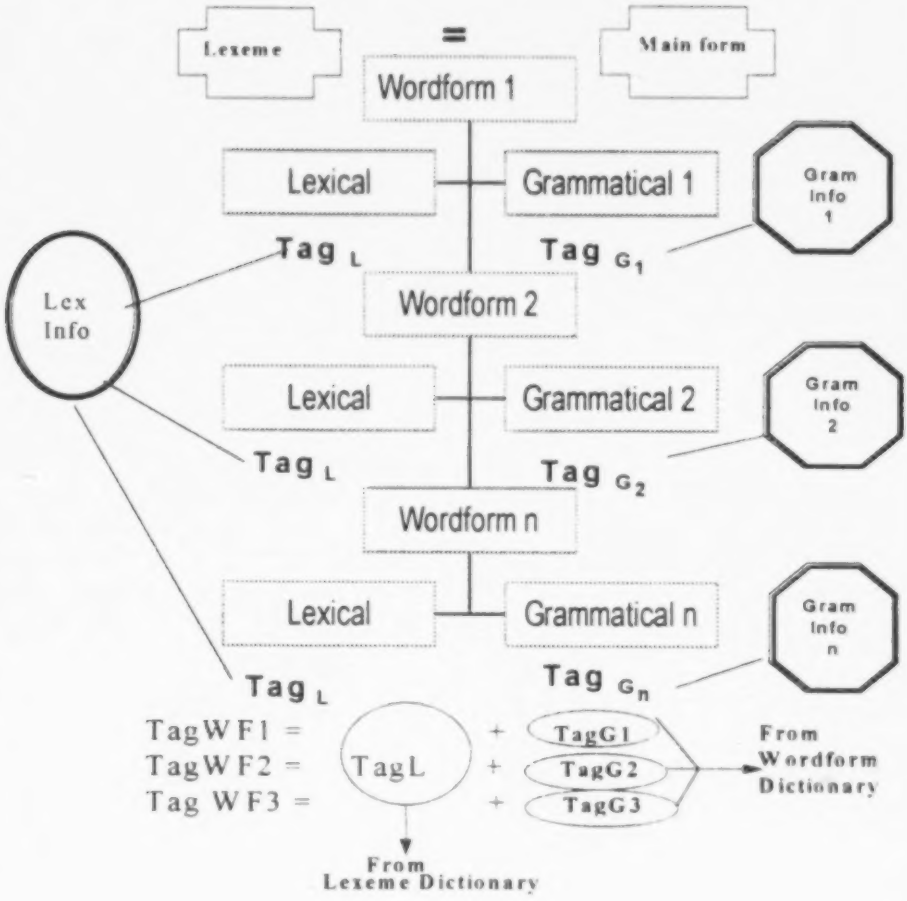


Figure 2. The constituents of tagging information for a highly inflected language

BEST COPY AVAILABLE

A *tagged* word has to include both types of knowledge (*lexical* and *grammatical*). An optimum organisation of tagging obviously has to provide for the separation of these two types of knowledge because one is obtained for every *class* and the other for every *representative* of the class. The proper distinction of these two types of knowledge (**TagL** and **TagGn**, respectively) is *conditio sine qua non* for every tagset description of a language which is a bit more inflected than English . We can see an unfortunate example of the blending of lexical and grammatical knowledge in the tagset of nine European languages in (1).

The two stages of tagging are – *lemmatisation* (transition from **WF** to **L** with extraction of **LexInfo**) and *grammatical analysis* (assignment of **GramInfo** to every **WF**).

The result of tagging (**TagT**) is a sequence of **TagW**. A **TagW** is a sequence representing **WF** as: **L, LexInfo** and **GramInfo**. The main menu of SUPERLINGUA performs the transition **WF–L** with the assignment of **LexInfo** and **GramInfo** to every unit from the **WF** dictionary (the latter is already constructed automatically during the preprocessing) (see Figure 3).

3.3 Optimisation and Acceleration of Manual Tagging

The lack of a morphological component does not automatically lead to the organisation of tagging as processing of *every* word in the textual sequence. The knowledge of the systematic organisation of an inflectional language allows us to accelerate that process.

1. The definition of **LexInfo** as class characteristics saves the repeated assignment of **LexInfo** to *all* class (paradigm) *members*. The more highly inflectional the language is, the greater the economy.
2. Some more time is saved with the help of the minimal morphological knowledge included in the system which contains the constant grammatical features (**GramInfo**) of the word form selected as a basic form in every POS class. The actual tagging of that word form is saved.
3. Also, the transition **WF – L** in the main menu conveniently includes the first part of the next disambiguating procedure – the disambiguation as an operation. Here the *rough disambiguation* is fulfilled, the definition of the lexemes for the homonymous wordforms presented in the processed text (not in the whole language). The context support helps us to correlate the list of the homonymous wordforms (**WFList**) to the list of their lexemes (**LList**) but preceding the tagging. Here, formally coinciding word forms are

attached to different lexemes after contextual help, their place in the text (and the correlation **WF-L**) remains undefined. This is the content of the *real disambiguation* performed on the real text during the next stage (and not on dictionaries of **WF** and **L** processed in the main menu).

Thus, in SUPERLINGUA, tagging of nonambiguous elements is carried out in the dictionary of word forms (**WfD**) and not in the linear text (which saves us *n-1* taggings where n is the number of appearances of the word form). There is no direct manual tagging of the all running words in SUPERLINGUA (see Figure 3).

Figure 3. Processed language units, their organisation, and operations in SUPERLINGUA

In this way, the *actual tagging* – processing of the *textual* sequence has to include only:

a) eventual error corrections – in the tagged text conveniently presented for editing, and
b) the second part of the disambiguating procedure – the distribution of the already tagged wordforms in **WfD** in the text (linking the wordforms as *dictionary units* to wordforms as *text units* – appearances).

Apart from the organisation of linguistic knowledge, acceleration of manual tagging can also be achieved through the very organisation of the processing operation. This includes: the division of text files into internal portions of 12KB which are seen by the user only in contextual help; the options allowing interruption of the processing of long files with the options *new* and *next* and their appropriate combination with *save* and *update*.

3.4 Additional Operations

Tagging is the final, deepest function of SUPERLINGUA. Other operations, which utilise the results of pre-processing and the transition **WF-L**, can extract almost all possible combinations of available knowledge and data organisation executed before or along with it. These are common and trivial operations such as *concordances* of various types, the construction production of *frequential* and *reverse vocabularies, string search*, standard *set operations* on vocabularies with *statistical data*, etc. Although they do not accelerate tagging directly, they are a useful secondary product of SUPERLINGUA and are manifestations of the above mentioned principle of extraction of all possible information from the available organisation of linguistic data (to give *the linguist all he wants...*)

3.5 Distribution of the System

The designers of the system would like to make their contribution to the fight against terminology abuse and misuse in which the combination of several standard sorting, searching, and string-matching procedures are presented as intelligent NLP achievements (terminology abuse is located on all levels of the linguistic hierarchy (from students' graduate theses to research projects). A well-known way to fight price-speculation is dumping sales. That is why the authors' intent is to distribute the system in the public domain for research.

References

Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicon and Corpora. A Common Proposal and Applications to European Languages. Draft version. EAG-LSG/IR-T4.6/CSG-T3.2. Project EAGLES, October 1994.

Seneca, De ira. 1970. Moral essays, vol. I. The Loeb Classical Library, London p.106

181

A Simple Czech and English Probabilistic Tagger: a Comparison

Barbora Hladká, Jan Hajič

Institute of Formal and Applied Linguistics
Malostranské nám. 25
118 00 Praha 1
Tel.: +42 2 21914 288
Fax: +42 2 21914 309
E-mail: {hladka, hajic}@ufal.mff.cuni.cz

1. Introduction

Highly inflectional languages like Czech pose a special problem for morphology disambiguation (which is usually called tagging). For example, the ending -u is not only highly ambiguous, but at the same time it carries a complex information: it corresponds, e.g., to genitive singular for inanimate nouns, or dative singular for animate nouns, or accusative singular for feminine nouns, or first person singular present tense active participle for certain verbs.

Given the success of statistical methods in different areas including text tagging we wanted to try them even for the Czech language one of the main features of which is a rich inflection displaying a high degree of ambiguity. Originally we expected that the result would be plain negative, getting not more than about two thirds of the tags correct. However, as we show later, we got better results than we had expected.

We used the same statistical approach to tag both the English text and Czech text. For English, we obtained results comparable with the results presented in [Brill 1993] (who uses different methods). For Czech, we obtained results which are less satisfying than those for English results.

2. Data Used

2.1 For Czech

For training, we used the corpus collected at the beginning of the 70ies in the Czechoslovak Academy of Sciences. The corpus was originally hand-tagged, including the lemmatization and syntactic tags. The complete size of the corpus is 600k tokens. We had to do some cleaning and conversion, as we were interested in the words and tags only.

2.2 For English

For training, we used Wall Street Journal [Marcus, Santorini, Marcinkiewicz 1993]. We had to change the format of WSJ to prepare it for our tagging software.

3. Tags

3.1 Czech tags

The original tag system (in the hand-tagged corpus) was too detailed to use it directly. We disregarded all the other information (lemmatization and syntactic tags) from the training data. We used the traditional division into the part of speech tagger classes. Each class contains many tags for each combination of morphological categories. For a description of the tags for the part of speech classes see Table 1. The first letter represents the tag for the part of speech class and it is followed by the morphological categories for the given class. We used special tags for sentence boundaries, punctuation and "unknown tag". We used 1171 different tags in our experiment for Czech. They were manually derived from the training corpus.

nouns	N	gender number case		
	abbreviation	Z		
adjectives	A	gender number case degree negation		
verbs	V			
	infinitive	T negation		
	transgressive	W number tense voice gender negation		
	common	person number voice tense mood gender negation		
pronouns	P			
	personal	P	person number case	
			3 gender number case	
	possessive	R	gender-of-the-possessive number-of-the-possessive	
			case person gender number	
	svůj	S	gender number case	
	se	E	case	
	others	D	gender number case negation	
adverbs	O			
conjunctions	S			
numbers	C			
prepositions	R			
interjections	F			
particles	K			

Table 1

For example:
NMS1 (noun, masculinum animate, singular, nominative)
NNP7 (noun, neuter, plural, instrumental)
VTA (verb, infinitive, affirmative)
V3SAPOMA (verb, 3rd person, singular, active, present tense, indicative, mas. anim., affirmative)
PP2P7 (personal pronoun, 2nd person, plural, instrumental)
AFP32N (adjective, femin. plural, dative, comparative, negative)

3.2 English tags

We used *The Penn Treebank* tagset which contains 36 Part-Of-Speech tags and 12 other tags (for punctuation and the currency symbol). A detailed description is available in [Santorini 1990].

4. The algorithms

We have used Merialdo's methods (described e.g., in [Merialdo 1992]). The tagging procedure selects a sequence of tags T for the sentence W:

$$\Phi : W \rightarrow T = \Phi(W).$$

In this case the optimal tagging procedure is

$$\Phi(W) = \underset{T}{\text{argmax}} Pr(T \mid W) = \underset{T}{\text{argmax}}\, Pr(T \mid W) * Pr(W) = \underset{T}{\text{argmax}} Pr(W \mid T) =$$

$$\underset{T}{\text{argmax}} Pr(W \mid T) * Pr(T)$$

Our implementation is based on generating the (W, T) pairs by a probabilistic model using approximations of probability distributions Pr(W | T) and Pr(T).

The Pr(T) is based on tag bigrams, and Pr(W | T) is approximated as the product of $Pr(w_i \mid t_i)$. The parameters have been estimated by the usual maximum likelihood training method, i. e. we approximated them as the relative frequencies found in the training data, smoothing them accordingly using the unigram frequencies and the uniform distribution.

5. The results

	Experiment for Czech	Experiment for English
corpus	**Czech hand-tagged**	**Wall Street Journal**
trainig data (tokens)	**621 015**	**1 287 749**
trainig data (words)	**72 445**	**51 433**
trainig data (tags)	**1 171**	**45**
training data (the average number of tags per token)	**3,65**	**2,34**
test data (tokens)	**1 294**	**1 294**
incorrect tags	**56**	**41**
tagging accuracy	**81,53%**	**96,83%**

To illustrate the results of our tagging procedures, we present here an example from the tagged test text. The cases of incorrect tag assignment are denoted by boldface letters.

tagged word | *hand-assigned tag* | *result of the tagging programme*

Czech test text

jménem | Rjménem | NNS7
úv | NZ | NZ
Ksč | NZ | NZ
pozdravil | V3SAMOMA | NZ
Davisovou | NFS4 | **NZ**
Pavel | NMS1 | NMS1
Auersperg | NMS1 | NMS1
W_SB | T_SB | T_SB
účastníci | NMP1 | NMP1
shromáždění | NNS2 | NNS2

English test text

In | IN | IN
the | DT | DT
lengthy | JJ | JJ
discussion | NN | NN
that | IN | **WDT**
followed | VBD | VBD
, | , | ,
Mr. | NNP | NNP
Buffett | NNP | NNP
said | VBD | VBD
: | : | :

6. Conclusion

The results, however they might seem negative compared to English, are still better than our original expectations. We would like to improve current approach by another simple measures. For example, the average number of tags per token will increase after a morphological analyser is added as the front end to the tagger (serving as the "supplier" of possible tags). We also plan to use trigrams instead of bigrams after we collect more data for Czech. Finally, certain tagset reductions be carried one, as the original tagset (even after the reductions mentioned above) is too detailed (in the sense that it distinguishes tags hardly distinguishable by human annotators). We are also working on independent predictions for certain grammatical categories and the lemma itself, but the final shape of the model has not yet been decided. This would mean to introduce constraints on possible combinations of morphological categories and take them into account when "assembling" the final tag.

References

Brill, E. 1993. "A Corpus Based Approach To Language Learning". Dissertation in Department of Computer and Information, Science, University of Pennsylvania.

Marcus, M. P., B. Santorini and M.A. Marcinkiewicz. 1993. "Building a large annotated corpus of English: the Penn Treebank". To appear in Computational Linguistics. (forthcoming)

Merialdo, B. 1992. "Tagging text with a probabilistic model". Computational Linguistics 20(2), 155–171.

Santorini, B. 1990. "Part of Speech tagging guidelines for the Penn Treebank Project". Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.

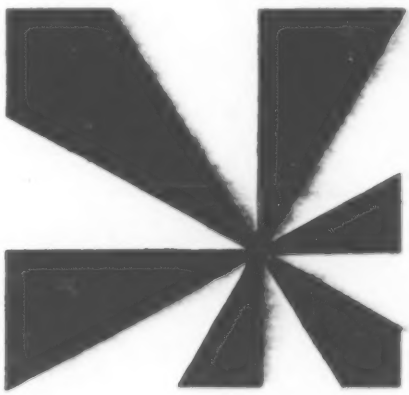
PRINTED IN HUNGARY

WHAT IS TELRI

The Concerted Action TRANS-EUROPEAN LANGUAGE RESOURCES INFRASTRUCTURE (TELRI), is a COPERNICUS project funded by the European Commission. TELRI has a duration of three years (1995–1997). It brings together 22 institutions of 17 European countries (Albania, Germany, Great Britain, Slovakia, Italy, Bulgaria, the Czech Republic, Sweden, Slovenia, Romania Estonia, France, the Netherlands, Latvia, Lithuania, Poland and Hungary).

TELRI will set up a permanent network of leading national language and language technology centres in the whole of Europe. It will pool existing language resources, corpora, machine-readable dictionaries and lexicons, lexical databases, and generic software tools for the creation, re-use, maintenance, validation, and exploitation of linguistic data. It will complement these repositories with newly created multilingual resources, offering a wide range of language data to the NPL community. TELRI will establish a platform where research and industry meet, exchange resources and engage in product-oriented cooperation.

Links have been established with language centres elsewhere in Europe, with relevant European organisations and ventures, and with focal language institutions in other parts of the world.



TELRI's WWW Document

Detailed information about TELRI and its activities is available through the World Wide Web (WWW) at the following URL:

http://www.ids-mannheim.de/telri/telri.html

FOR INFORMATION:

Inquiries about TELRI may be addressed to:

Dr. Wolfgang Teubert
Institut für deutsche Sprache
P.O. Box: 101621
68016 Mannheim, Germany
Phone: +49 621 1581 437
Fax: +49 621 1581 415
Email: telri@ids-mannheim.de

BEST COPY AVAILABLE